Balancing The Perception of Cheating Detection, Privacy and Fairness: A Mixed-Methods Study of Visual Data Obfuscation in Remote Proctoring

Suvadeep Mukherjee
suvadeep.mukherjee@uni.lu
University of Luxembourg
Luxembourg

Verena Distler*
verena.distler@aalto.fi
University of the Bundeswehr Munich
Germany

Gabriele Lenzini
gabriele.lenzini@uni.lu
Interdisciplinary Centre for Security Reliability and Trust (SnT), University of Luxembourg
Luxembourg

Pedro Cardoso-Leite
pedro.cardosoleite@uni.lu
University of Luxembourg
Luxembourg

Abstract

Remote proctoring technology, a cheating-preventive measure, often raises privacy and fairness concerns that may affect test-takers' experiences and the validity of test results. Our study explores how selectively obfuscating information in video recordings can protect test-takers' privacy while ensuring effective and fair cheating detection. Interviews with experts (N=9) identified four key video regions indicative of potential cheating behaviors: the test-taker's face, body, background and the presence of individuals in the background. Experts recommended specific obfuscation methods for each region based on privacy significance and cheating behavior frequency, ranging from conventional blurring to advanced methods like replacement with deepfake, 3D avatars and silhouetting. We then conducted a vignette experiment with potential test-takers (N=259, non-experts) to evaluate their perceptions of cheating detection, visual privacy and fairness, using descriptions and examples of still images for each expert-recommended combination of video regions and obfuscation methods. Our results indicate that the effectiveness of obfuscation methods varies by region. Tailoring remote proctoring with region-specific advanced obfuscation methods can improve the perceptions of privacy and fairness compared to the conventional methods, though it may decrease perceived information sufficiency for detecting cheating. However, non-experts preferred conventional blurring for videos they were more willing to share, highlighting a gap between the perceived effectiveness of the advanced obfuscation methods and their practical acceptance. This study contributes to the field of user-centered privacy by suggesting promising directions to address current remote proctoring challenges and guiding future research.

CCS Concepts

• **Human and societal aspects of security and privacy**; • **Applied computing** → *E-learning*; *Distance learning*;

Keywords

Remote Proctoring, Cheating, Obfuscation, Privacy-Utility, Fairness, UX, Willingness to Share

ACM Reference Format:
Suvadeep Mukherjee, Verena Distler, Gabriele Lenzini, and Pedro Cardoso-Leite. 2024. Balancing The Perception of Cheating Detection, Privacy and Fairness: A Mixed-Methods Study of Visual Data Obfuscation in Remote Proctoring. In *The 2024 European Symposium on Usable Security (EuroUSEC 2024), September 30-October 1, 2024, Karlstad, Sweden.* ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3688459.3688474

1 Introduction

In remote high-stakes testing, aimed at cost-effective competence assessment, cheating prevention is often done via video recording, known as remote proctoring. It comes in various forms, including human-involved (remote live invigilation or recorded sessions with post-test verification) and AI-driven (solely operated by algorithms) [19]. Institutions often opt for a hybrid approach, combining algorithms to flag suspicious behaviors during tests with subsequent review of recordings by professionals [27], referred to as 'reviewers' in this paper. Although remote proctoring is often mandated by institutions [45] as anti-cheating measures, it can adversely affect test-takers' experiences due to significant privacy concerns [1, 7, 32]. This is because the video recordings may inadvertently disclose sensitive information about individuals (e.g., their family details, religious beliefs, disabilities or gender identities [6, 23]). Additionally, test-takers may feel uncertain about how their videos are observed by the reviewers [6, 46]. For instance, AI algorithms in proctoring software, often trained on biased datasets, might unjustly flag cheating incidents. When presented to reviewers, these biases can influence their judgments, potentially leading to unfair accusations of cheating [2, 57].

This study aims to address these issues by focusing on hiding privacy-sensitive video content from reviewers to enhance test-takers' privacy perception. Leveraging recent advancements

*Current affiliation: Aalto University, Finland

EuroUSEC 2024, September 30-October 1, 2024, Karlstad, Sweden

ACM ISBN 979-8-4007-1796-3/24/09
https://doi.org/10.1145/3688459.3688474

in computer vision, the study explores the scope of different obfuscation methods used in various video conferencing tools (e.g., Zoom, Microsoft Teams) and social media platforms (e.g., Instagram, Snapchat) to alter appearances while preserving motion in videos. A range of methods, including blurring [34], pixelation [18], masking [39], inpainting [28] and augmented reality 3D animations [14], can be used in live-streaming scenarios [17]. Additionally, more advanced methods like deepfake [48], which synthesizes realistic visual content, can be implemented using sophisticated video editing tools. This study also examines how obfuscation methods impact test-takers' perceptions of potential discrimination by reviewers, as they can hide attributes like test-takers' ethnicity, age or gender [40], potentially preventing unfair accusations [57]. While these methods can protect privacy and reduce discrimination, they can also hide visual data crucial for identifying cheating behaviors like movements in head, mouth or eye gaze [38]. This perceived reduction in cheating detection is crucial as it may influence test-takers' likelihood to cheat, regardless of the actual detection capabilities [13].

Hence, our research aims to: *(1)* identify the most relevant obfuscation methods that can be applied in remote proctored videos, and *(2)* investigate how potential test-takers perceive the effectiveness of those identified methods in terms of cheating information sufficiency, privacy protection and fairness in cheating detection. Additionally, to assess whether these methods improve user experience (UX), we measured the willingness of potential test-takers to share obfuscated videos with reviewers as an indicator of their trust and attitude toward these methods [10]. To the best of our knowledge, no prior research has investigated these three crucial aspects collectively when obfuscating video content in remote proctoring to enhance test-takers' experiences. For this purpose, we conducted a two-part study using a mixed-methods approach, combining a qualitative and quantitative study.

The first part, addressing the following research questions, involved semi-structured interviews with experts to explore the significance of obfuscating specific regions (face, body, background and other individuals) in a video.

RQ1: What specific visual information do experts believe different video regions offer for cheating detection?

RQ2: What obfuscation methods do experts believe should be relevant in each region to address test-takers' privacy concerns while ensuring sufficient information for cheating detection and avoiding unfair accusations?

RQ3: What recommendations do experts provide for applying the identified obfuscation methods on the video?

In the second part, to address the following research questions, we evaluated the effectiveness of the expert-recommended region-specific obfuscation methods with non-expert potential test-takers in a vignette experiment. To simplify the user-testing process, we used simulated images as a proxy for remote proctored videos with each of the obfuscation methods applied and asked participants to visualize them as visual scenes in the video recordings.

RQ4: What are the effects of the region-specific obfuscation methods on test-takers' perceptions of cheating information sufficiency, privacy protection and fairness in cheating detection?

RQ5: What are the effects of the region-specific obfuscation methods on these perceptions combined?

RQ6: What are the effects of the region-specific obfuscation methods on test-takers' willingness to share videos with unknown reviewers if obfuscated with those methods?

1.1 Contributions

- Our study provides a research direction toward improving test-taking experiences in remote proctoring by hiding sensitive visual data in recorded videos. The aim is to select obfuscation methods pragmatically not only for their capacity to balance perceptions of privacy protection and fair cheating detection but also for their potential to foster trust among test-takers, ensuring exam integrity while preserving privacy.
- Our study proposes that obfuscation methods should be tailored to regions (face, body, background and other individuals) in a remote proctored video based on their privacy significance and frequency of cheating behavior.
- We explored different obfuscation methods with distinct functions, eliciting different qualitative perceptions. Additionally, we highlighted fairness concerns to guide researchers considering their implementation in videos.

2 Related Work

2.1 Cheating Detection in Remote Proctoring

In high-stakes remote testing, ensuring integrity is crucial for detecting cheating, such as unauthorized resource use or external assistance. Remote proctoring, a common cheating-preventive method, typically involves three setups, each with its own challenges [19]. The first involves live monitoring by remote invigilators, who intervene immediately for exam rule violations. This is resource-intensive and lacks test-taking flexibility. The second records the entire test session for later professional review to identify cheating behaviors by watching the entire video, but it's time-consuming. The third relies on AI/ML algorithms alerting test-takers if a suspicious behavior is detected during test-taking, but raises concerns due to the nature of the datasets the algorithms are trained with. Test-organizing institutes often use a hybrid approach [27] to address these shortcomings, ensuring cost-effectiveness. The process includes recording the test-taking session, with algorithms generating reports of detected cheating behavior [38] and sending these reports to reviewers afterward, enabling them to access relevant timestamps, hence streamlining the process [46]. Various suspicious events (e.g., test-takers being absent from the frame, the presence of other individuals, different individuals taking the test and students disabling the webcam) have been documented as potential cheating behaviors [57]. Our study focuses on the hybrid setup, promising for obfuscation applications and streamlining cheating verification.

2.2 Privacy Concerns with Video Recordings in Remote Proctoring

In recent times, concerns over privacy perception have surged due to the abundance of detailed visual content captured by video recording devices. From public surveillance cameras capturing individuals without consent [3, 52] to the use of remote proctoring in high-stakes testing, mandated by institutional obligations [45],

unintended recording of sensitive information like test-takers' family members [22], religious affiliations [23], disabilities, or gender orientations [6] may not be considered necessary for ensuring test integrity. Despite encryption and access control measures, increasing online data breaches have heightened user concerns about personal data privacy [9, 20], potentially leading to resistance [45] and legal actions [31, 53]. Therefore, privacy-enhancing measures should be perceived as robust strategies by testing stakeholders, particularly by test-takers, as they can influence test-taking experiences [1] and subsequent adoption of the practice [49]. These measures should reassure test-takers that their videos are not inappropriately monitored or pose harm if shared with third parties [6], preventing unnecessary privacy concerns and distress [1].

2.3 Fairness Concerns with Cheating Detection Outcome in Remote Proctoring

In hybrid remote proctoring, where reviewers consult algorithm-generated reports of detected suspicious behaviors before verifying corresponding video timestamps, risks might arise in terms of fair cheating detection. For instance, if reviewers prioritize algorithm reliance over objective judgment, unfair cheating accusations can occur due to algorithmic biases from skewed training datasets [2, 57]. Studies [57] indicate that algorithms tend to flag individuals with darker skin tones more frequently for cheating allegations, particularly affecting females with dark skin tones. Moreover, algorithms may flag unconventional student actions during tests (e.g., unusual head movements, muttering, looking to the side, leaning on hands, wiping faces, drinking water or approaching the screen closely) leading to increased false positives [27, 57]. Additionally, reviewers' personal biases, influenced by factors like test-takers' ethnicity, skin tone and gender, can also result in unfair accusations [15]. Given these limitations, a privacy-preserving measure should provide assurance to test-takers that such biases will not impact the fairness of cheating detection outcomes.

2.4 Obfuscating Video Contents in Remote Proctoring

Privacy protection in video-based applications often involves obscuring identifiable video content. Recent advancements in AI and computer vision allow for real-time alteration of appearances in video chats or post-editing of recorded videos using visual obfuscation methods like facial filters and dynamic augmented reality animations, widely popular on social media (e.g., Snapchat, Instagram etc.) and increasingly integrated into video conferencing platforms (e.g., Zoom, Microsoft Teams). Some of the prevalent methods include conventional and lightweight methods like *blurring* [34] and *pixelation* [18] that alter a region by recalculating neighboring pixel values; *masking* with a solid box [39]; *inpainting* to fill missing parts [28]; to more advanced methods like *deepfake* technology for synthesizing video content [48] or replacing individuals with *3D cartoon avatars* [14] and *silhouette-like* figures by contour masking [29]. While concealing identifiable video content can improve the user experience by preserving privacy perception [21], the selection of methods often relies on the objective evaluations of various factors within specific contexts [11, 39].

One key consideration is the region of the video where obfuscation is applied; for example, studies show that hiding the entire body enhances privacy perception more than just hiding the face [5], and masking is more effective than blurring in that task [29]. However, it can also lead to the loss of crucial details relevant to the context; for instance, masking of test-takers in remote proctored videos can obscure vital behavioral cues like body movement or eye gaze, underscoring the importance of balancing privacy and utility [39]. Obfuscation methods also vary in their computational demands [59]; for instance, applying deepfake to synthesize the test-taker's features during video editing could demand high processing power, affecting test institutions' budgets. Furthermore, evaluating methods often involves assessing their effectiveness against not only human observation (perceptual obfuscation) but also adversarial algorithmic attacks aimed at reversing obfuscation (machine obfuscation) [39], such as the potential identity reversibility seen in blurring and pixelation [25]. Given the various factors influencing such evaluations, there is a need to pragmatically select obfuscation methods, which is explored in Section 4 followed by user evaluation in Section 5.

2.5 User Experience in Remote Proctoring

In recent years, assessing user experience (UX) has gained popularity to measure technology's impact on users by going beyond usability (i.e., ease of use, efficiency) and including emotional responses, trust and beliefs resulting from user interaction with a digital product [16]. Technologies requiring users to share personal information often face adoption challenges [49] due to concerns about transparency in data usage and potential misuse [24], as well as the risk of discrimination based on identifiable data, impacting user trusts [54]. In remote proctoring, where test-takers share test-session videos under obligatory conditions amid resistance due to poor test-taking experiences [6, 45], assessing their trust in obfuscation methods applied to video content is crucial for UX, alongside their objective evaluations for privacy safeguards and fair cheating detection. Measuring trust through their willingness to share video [49] under applied obfuscation can guide us assessing adoption and hence standardizing video recording for test integrity.

3 Study Design

This study, divided into two parts, aims to identify the most promising obfuscation methods for remote proctoring video recordings and to determine how these methods, by hiding sensitive and potentially discriminatory data in the videos, affect test-takers' perception of privacy protection and fairness in cheating detection. Simultaneously, the obfuscation methods must not compromise cheating information sufficiency, as it can influence their test-taking experience and likelihood of cheating, assuming they believe that cheating actions can be identified. In Part 1, experts recommended obfuscation methods for hiding specific video regions, known as Regions of Interest (ROIs). Part 2 evaluated these recommended methods with potential test-takers, assessing their perceptions of privacy, fairness and information sufficiency, as well as their willingness to adopt each method for remote proctoring.

4 Part 1: Identifying Suitable Obfuscation Methods in Remote Proctoring

This section gathers insights from expert interviews regarding factors to consider when identifying obfuscation methods relevant for hiding specific video regions. It also aims to propose promising remote proctoring pipeline, such as if obfuscation should occur in real time or post-test and the associated tasks of different test stakeholders.

4.1 Recruitment, Interview Protocol and Analysis

For interview purpose, we recruited nine experts based in the USA and Europe through email within the researchers' network (details are provided in Appendix Table 10). They comprised three professionals with expertise in remote proctoring within universities, four specialists engaged in computer vision research in the industry and academia and two researchers specialized in usable privacy, also serving as professors in universities. Each expert engaged in a session lasting 1-1.5 hours and received €40 for their time. The sessions were conducted remotely in a semi-structured format.

The first author facilitated expert engagement via an online collaboration platform, 'miro.com'. Interview materials are provided in Appendix Figure 5. Initially, experts identified visual cues indicating potential cheating behaviors in different regions (faces, bodies, backgrounds) of a visual scene by examining a simulated image of front-facing test-takers with visible body parts. They rated the importance of these cues on a scale from 1 (least) to 10 (most) for cheating detection. Next, they evaluated five pre-prepared images with various obfuscation methods applied to faces (blurred, pixelated, masked, deepfaked avatarized), placing them on a 2D-MAP with the x-axis representing privacy protection and the y-axis as cheating detection difficulty, while elaborating on their decisions. This process was repeated by asking experts to visualize similar obfuscation methods applied to other regions. Experts also suggested measures to mitigate potential biases by hiding potentially discriminatory attributes to ensure fair judgments in cheating detection. Finally, experts proposed a viable solution for a cost-effective obfuscation pipeline for remote proctoring.

To extract insights, the first author analyzed both task outcomes on 'miro.com' and related discussions. The interviews were transcribed and analyzed using deductive content analysis in MAXQDA (v.2024). Initial coding involved five main categories: identifying cheating behaviors for each region, assessing the importance of visual cues for detecting cheating behaviors, evaluating the advantages and disadvantages of obfuscation methods in each region, proposing measures to address fairness concerns and discussing the obfuscation pipeline. The transcripts were thoroughly reviewed and the relevant content were highlighted, coded and categorized accordingly. The codebook is provided in Appendix B.3.

4.2 Results

4.2.1 Significance of Regions of Interest (ROI) for Cheating Detection in Visual Scenes. The focus was on distinguishing possible cheating instances linked to foreground (facial and body regions) and background (stationary and moving elements like individuals appearing behind during tests) regions or ROIs.

Most experts, particularly those with experience in remote proctoring, identified that the most common and frequent cheating activities originate from the face region, with indicators such as mouth, head and eye movement being cited as the most frequent. This underscores the reviewers' need for significant attention in this area. Almost all experts stressed that potential cheating instances within an ROI shouldn't be viewed in isolation, as cheating behavior is multifaceted and may involve various cues. For example, body movement accompanied by a change in eye gaze direction or a synchronized mouth movement between a test-taker and a person in the background might indicate suspicious behavior. All identified cheating instances to each ROI, along with potential associations with other ROIs, are presented in Table 1. Beyond test-takers' face and body, the background region may also serve to detect cheating, for example, revealing unpermitted resources or interactions with other people. Remote testing guidelines often advise on suitable test-taking locations to prevent such disclosures, though enforcement may vary given the diverse living situations of test-takers.

Table 1: List of potential cheating occurrences linked to regions of interest (ROI)

Concerned ROI	Cues for suspicious behavior	Cheating instances	Related ROIs
Face	Mouth movement	Discussing answers using an earphone; asking for answers from someone present	People in background
	Eye movement	Looking away from the screen; interacting with someone in the room	
	Head movement	Interacting with someone in the room; lowering head to the desk, possibly using unauthorized materials	
Body	Body pose	Moving away from the screen; interacting with someone present	People in background
	Hand movement	Using unauthorized materials (e.g., smartphone); interacting with objects like books	
	Shoulder movement	Shifting shoulders to engage in cheating activities such as using phones, books, etc.	
Background	Presence of camera	Live-feeding computer screen for remote question dissemination	
	Visible cheat notes	Test-taker writes cryptic answers on posters or walls	
People in background	Mouth movement	Talking to the test-taker during the exam	Face, body
	Body movement	Approaching test-takers closely to assist or view the computer screen	

4.2.2 Obfuscation Methods to Balance Privacy Protection and Cheating Detection. Following the identification of potential cheating instances for each ROI, experts evaluated common obfuscation methods (e.g., blurring, pixelation, masking, 3D avatar representation and deepfake), focusing primarily on privacy significance and the frequency of cheating behavior in each ROI. Additionally, factors such as their effectiveness in improving privacy protection, retaining cheating information post-obfuscation, scope of reversibility by de-obfuscating algorithms and practical challenges like computational demands when preserving motion for faces, bodies etc. [39], were also taken into account.

Experts observed that once test-takers are authenticated before the test, facial and body features become less crucial for review, allowing for their obfuscation during video editing, provided that motion from these areas, such as eye and mouth movement or hand

gestures, is adequately preserved. Conventional obfuscation like blurring and pixelation were considered moderately effective for privacy protection and cheating information preservation across all ROIs, but masking was deemed unsuitable for the face and body regions due to eliminating crucial indicators of suspicious behavior and inability to preserve motion. For the face region, deepfake and 3D avatar replacement were preferred over other methods due to their superior privacy protection (in terms of both perceptual and identity irreversibility [25, 39, 48, 58]) and ability to retain most cheating information if motion is preserved. However, computer vision experts warned that they can be computationally more intensive than conventional obfuscation methods, raising concerns over institutional support and budget for video editing. Deepfaked face was somewhat preferred over 3D avatar replacement because the former offers more realistic identity replacement by providing more texture information [51]. For the body region, a straightforward approach like outfit replacement can be effective, especially if motion, including hand and shoulder movement and changes in body poses, is well preserved. For that purpose, blurring, pixelation and deepfake were rated by experts as viable options. To obfuscate infrequent instances like people appearing in the background, a variant of masking, such as silhouette-like figure replacement based on region contour, was suggested, as it captures suspicious body behavior while overlooking individual indicators like eye and mouth movements. Alternatively, a motion-preserving full-body 3D avatar could also be effective. To obfuscate the background, blurring, pixelation or replacing it with a picture can achieve the goal of concealing stationary objects. Table 2 summarizes the expert recommendations for obfuscation methods in each ROI. Note that, we opted for blurring over pixelation for our user evaluation phase in Section 5 because of their similar effectiveness [26], thus avoiding redundancy in user assessment.

Table 2: Relevant region-specific obfuscation methods as recommended by experts

Region of interest	Blurring	Silhouette	Deepfake	3D avatar
Face	✓	X	✓	✓
Body	✓	X	✓	X
Background	✓	X	✓	X
People in background	✓	✓	X	✓

4.2.3 Addressing Potential Fairness Concerns in Cheating Detection. Obfuscating face, body and other ROIs to enhance privacy protection may simultaneously suppress discriminatory attributes, potentially mitigating unfair judgments influenced by reviewers' personal biases. However, concerns about fairness may still arise regarding other discriminatory cues such as skin tone, ethnicity and gender of test-takers [15]. Privacy experts raised concerns about the possible interdependence between ethnicity and skin tone [57], suggesting that altering one without the other might not ensure fairness. Mixed opinions emerged regarding changing the skin tone when replacing a face with a different identity, prompting user-testing with altered skin tones. Caution was also advised to ensure consistency by extending alterations to other visible skin areas (e.g., neck, hands). Moreover, replacing all faces with a single ethnicity though may reduce discrimination, caution is needed regarding

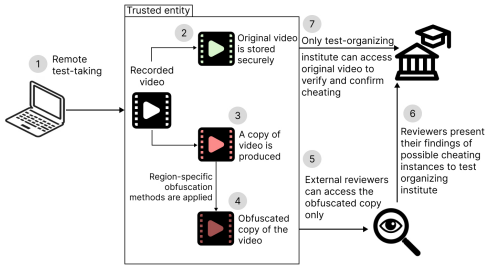

Figure 1: Expert recommended pipeline for video obfuscation

factors like ethnic attire or ornaments that could still reveal ethnicity. Addressing gender bias, including cues like hair textures and length, also requires careful consideration.

4.2.4 Practical Obfuscation Pipeline for Remote Proctored Videos. Considering the practical limitations of obfuscation, nearly all experts stressed that relying solely on obfuscated videos for detecting cheating could pose significant challenges during disputes over cheating allegations. To tackle this, they suggested securely storing an unprocessed copy of the recorded video by a trusted entity. Region-specific obfuscation should then be applied to the original video there, with access granted for review purposes. The unprocessed video should be restricted to the test-organizing institutions, enabling them to make final decisions [56]. The visual representation of the pipeline is depicted in Figure 1.

5 Part 2: Non-expert User Evaluation of Obfuscation Methods in Remote Proctoring

5.1 Methodology

5.1.1 Experimental Design. We conducted a vignette experiment with 259 non-expert potential test-takers to evaluate expert recommended obfuscation methods on perceptions of privacy protection, fairness, cheating information sufficiency and user experience (UX). We used a 4x4 within-subject design with four obfuscation methods (blurring, silhouette, deepfake and 3D avatar) across four ROIs: face, body, background and people in the background. However, we only tested obfuscation methods that are specific to an ROI as recommended by experts (see Table 2). Participants first received a brief introduction to the current challenges of video recordings followed by our research objectives, an example stimulus demonstrating obfuscation, and task instructions. The experimental design is depicted in Figure 2. Each ROI was presented with two gender variations, showing original stimuli followed by region-specific obfuscations alongside the originals, and participants completed various questionnaires (details are in Section 5.1.3) accordingly.

In the survey, we used still images as test stimuli instead of video for simplicity. Participants were instructed to visualize the image as a visual scene in a video recording. This approach was chosen for several reasons. Firstly, research [8, 41] indicates that static images or frames, which are sequenced to create videos, can capture essential aspects influencing user evaluations, even though they may not fully represent the complexities and nuances of dynamic video content. Additionally, static images might allow for better systematic control over variables in a visual scene compared to

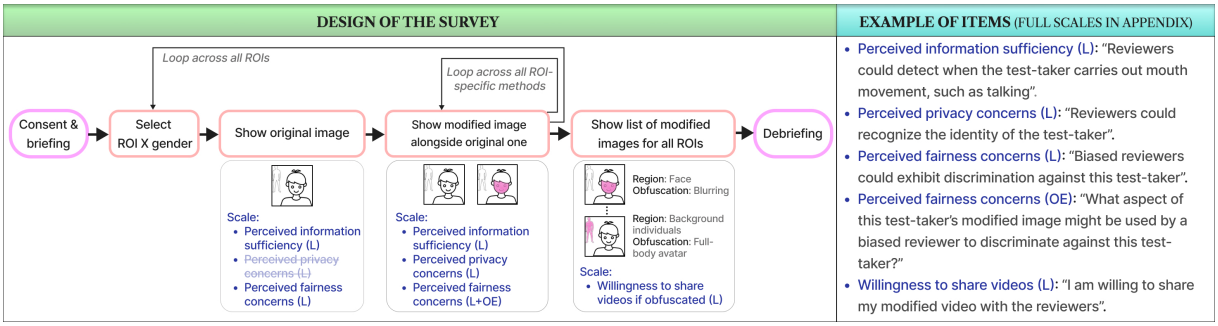

Figure 2: The survey presents an image as a visual scene, followed by hiding a region using various obfuscation methods. For instance, the test-taker's face above is highlighted to indicate where obfuscation is applied. Participants then provided their opinions using Likert scale (L) or open-ended (OE) items, with example items shown on the right side. Genders varied over the ROI in focus: for face, body and background obfuscation, test-takers' genders were varied; for background people, the genders of those people were varied

video-based stimuli [36], where a subtle change in context and how participants watch them (skipping or watching in entirety) can introduce variability in judgments [55].

5.1.2 Creation of Stimuli. The objective was to create a set of original images and apply obfuscation methods to them.

Creation of original images: The first step involved generating original images of test-takers in simulated webcam-captured scenarios, ensuring equal gender representation, featuring diverse skin tones and diverse test-taking environments such as personal rooms or office spaces. Faces, aged between 18-40 years, sourced from *thispersondoesnotexist.com*, were chosen for their realistic yet non-existent appearance, while Adobe Photoshop's (ver. 2024) Generative Fill feature was used to create other elements like the body and background. Faces from various ethnicities, including black, white, Arabic and Latin were included, resulting in two gender variations for each of the four ROIs, totaling 8 original images (refer to Figure 3), with ethnicities randomly distributed.

Creation of modified images: The second step involved creating an augmented set of stimuli by obfuscating the original images, following expert recommendations for region-specific obfuscation (see Table 2). This process, varied over two genders, resulted in 20 modified images. Additionally, two variations of deepfaked faces were created to evaluate potential fairness concerns regarding skin tone, resulting in a total of 22 images (refer to Figure 3). Next, we detail the various obfuscation methods employed.

1. ***Blurring***: We applied Gaussian blur with varying radii to balance cheating information sufficiency and privacy. Radii of 14 pixels for the face, 50 pixels for the body, 20 pixels for the background and 40 pixels for people in the background were found to strike this balance in pilot testing (see Appendix A.2). *2.* ***Deepfake***: To address potential gender bias in facial obfuscation, we opted for a gender-neutral reference face created with the FaceMaker application [44]. Using open-source deepfake tools [42], we swapped faces while maintaining original facial expressions' fidelity. To evaluate fairness [47], each deepfaked face had two versions: one preserved the original skin tone, while the other was altered to the category 3 in the Fitzpatrick scale [12], a widely accepted skin color standard. Additionally, we dressed test-takers uniformly in formal attire, adjusted slightly for each gender, and replaced backgrounds with freely available room pictures. *3.* ***3D avatar***: In contrast to realistic deepfake identities, cartoonized avatars provide less detailed visual information [51]. Using the FaceMaker application [44], we generated gender-neutral cartoonized 3D face avatars. These avatars replaced the original faces, with attempts made to align them with the facial pose and expressions using Photoshop. Similarly, for individuals in the background, the visible portions were altered to 3D full-body avatars, mimicking the subjects' facial and body expressions, using the Generative Fill feature of Photoshop. *4.* ***Silhouette***: We used a 2D silhouette representation only for the individuals in the background, filled with black color.

5.1.3 Measurements. We developed a questionnaire to assess test-takers' perceptions of *a) information sufficiency for cheating detection, b) privacy concerns, c) fairness concerns and d) willingness to share video recordings with reviewers* for each combination of regions and expert-recommended obfuscation methods. The items are described below with examples, and the full questionnaire can be found in the Appendix Table 5. Most items used a 7-point Likert scale ("strongly disagree" to "strongly agree") [4]. For perceived fairness concerns, we also included an open-ended question.

(a) Perceived Information Sufficiency: This refers to test-takers' belief that the obfuscated video contains enough information to detect cheating. This perception, whether accurate or not, can influence their willingness to cheat. Items were based on expert interview findings on different cheating instances in various regions (see Table 1). For example, we asked if test-takers thought, *"Reviewers could detect when the test-taker carries out mouth movement, such as talking"* for the obfuscated face region. ***(b) Perceived Privacy Concerns***: This refers to test-takers' belief in how well the obfuscation can suppress identifiable information of an individual. Items focused on the test-taker's perception of the reviewer's ability to identify obfuscated regions [29]. For example, we asked if they thought, *"Reviewers could recognize the outfits of the test-taker"* for the obfuscated body region. ***(c) Perceived Fairness Concerns***: This refers to test-takers' belief in whether obfuscation can effectively hide potentially discriminatory information that might lead biased reviewers to make false cheating allegations. We asked if

Figure 3: Visual scenes were created for our vignette experiment, manipulating four regions (face, body, background and people in the background) with relevant obfuscation methods. Both male and female subjects are represented, resulting in 8 original visual scenes. When deepfake was applied to the facial region, the questionnaire assessed both the original and changed skin tones of the subjects

they thought, *"Biased reviewers could exhibit discrimination against the test-taker"* for each obfuscated region. We also asked an open-ended question to assess the potential bias post-obfuscation. ***(d) Willingness to Share Videos***: Finally, to measure UX and preferences for sharing visual data [28, 29], we included items to assess test-takers' willingness to share videos with a reviewer for each combination of regions and obfuscation methods, such as *"I am willing to share my modified video with the reviewers"*.

In the survey, non-expert participants were shown 30 images: 8 original and 22 modified, which were prepared in Section 5.1.2. According to the study design in Figure 2, participants were sequentially presented with original images followed by their modified versions for each of the four regions. For the modified images, participants rated information sufficiency for cheating detection, privacy concerns and fairness concerns. Privacy concern items were skipped only for the original images since these were designed to assess the reviewers' ability to recognize specific ROIs, which we assumed could be identified without obfuscation. After these steps, participants rated their willingness to share videos for all combinations of regions and obfuscation methods. The complete survey is available in Section A.1 in the Appendix.

5.1.4 Recruitment and Ethical Considerations. To evaluate the expert recommended region-specific obfuscation methods, we conducted user testing with 259 UK-based remote participants recruited via Prolific, a reliable crowd-working platform known for providing high-quality data [35]. This sample size ensured robust statistical inference, crucial for guiding potential applications in remote proctored videos. Adult participants aged 18 to 60 were considered eligible and were then directed to the survey. Data collection took place in January 2024. On average, participants took 25 minutes to complete the survey. The sample was non-representative, with 51.3% female, 46.7% male, and 2% non-binary participants. On average, participants were 32 years old (SD=10) with diverse educational backgrounds and around 67% had a university degree. Approximately 37% had taken at least one remote proctored test in the last three years.

At the beginning of the survey, participants were given a digital informed consent form and a study information sheet. We followed GDPR practices and informed participants about data collection, storage and opt-out opportunities. Following completion, participants received a written debriefing explaining the study's purpose and the creation of images for research purposes. Each participant received compensation as per Prolific's hourly pay policy. The university's ethics committee reviewed and approved our research project.

5.1.5 Data Analysis. In the following section, we address three research questions: *(1) RQ4* investigates the effect of region-specific obfuscation methods on perceptions of information sufficiency,

privacy and fairness; *(2) RQ5* examines their effect on the overall perceptions; and *(3) RQ6* investigates their effect on participants' willingness to share videos with unknown reviewers if obfuscated. Following data collection, we prepared the dataset for dependent variables in three steps. First, we assigned a maximum value of 7 uniformly to all participants for the skipped privacy concerns items when the original images were shown (as explained in Section 5.1.3). We then reversed the Likert scale responses for privacy and fairness concerns to align with our research objectives, considering a rating of 1 as 7 and vice versa, with higher scores indicating greater privacy protection and fairness. Next, we computed the mean values of the dependent variables if multiple items were asked within each ROI. Finally, we averaged the ratings across the two image variations shown per ROI. Table 3 presents the data prepared for analyzing the dependent variables.

Statistical Analysis: To evaluate the impact of obfuscation methods on each dependent variable across ROIs, we performed linear mixed-effects models (LMEMs). LMEMs are particularly suited for this analysis because they can account for the variability in participants' responses when multiple ratings are collected from the same participant. Although the Likert scale ratings are ordinal in nature, LMEMs could still be performed for several reasons: (a) the 7-point Likert scale that we used is typically regarded as having approximately equal intervals between levels [4]; and (b) Likert scales with five or more response categories are commonly treated as continuous in statistical analyses [33]. However, potential limitations persist when treating ordinal data as continuous in parametric methods like LMEMs. This approach can introduce biases in parameter estimation and lead to interpretation challenges, including false alarms, loss of power and incorrect ordering of means [30].

In our LMEMs, the independent variable was the obfuscation method applied to each ROI. The model used 'no obfuscation' as the baseline for the dependent variables i.e. perceived information sufficiency and perceived fairness, which were rated without obfuscation. For the remaining dependent variables, 'blurring' was used as the baseline, due to its conventional use. Initial exploratory analyses indicated that the residuals were neither normally distributed nor homoscedastic. To address these violations of LMEM assumptions, we employed robust standard errors. Post-model diagnostics showed that the results with and without robust standard errors were consistent (see Appendix Table 9), suggesting that our findings were not significantly affected by these assumption violations.

After fitting the LMEMs, we conducted post hoc pairwise comparisons between obfuscation methods, with a focus on estimating effect sizes by measuring marginal mean differences (MMD). All regression results are presented in Table 7 in the Appendix, and the detailed pairwise comparisons are provided in Table 4. All statistical analyses were conducted using STATA v18.

5.2 Results

5.2.1 Effects of Region-specific Obfuscation Methods on Perceived Information Sufficiency, Perceived Privacy and Perceived Fairness. Based on the prepared data in Table 3, PANEL 1 in Figure 4 illustrates the average ratings for perceptions of information sufficiency, privacy and fairness relative to 'no obfuscation' condition for each combination of ROIs and obfuscation methods. These perceptions

Table 3: Mean values with std. deviations of all dependent variables for each combination of ROIs and obfuscation methods. The means of perceived privacy protection in 'no obfuscation' were considered 1.00 for analysis purposes, as explained in Section 5.1.5

Region of interest (ROI)	Obfuscation methods	Perceived info. sufficiency	Perceived privacy protection	Perceived fairness	Composite scores	Willingness to share videos
Face of test-taker	No obfuscation	6.14 (0.85)	-	2.25 (1.13)	3.13 (0.39)	-
	Blurring	4.57 (1.32)	4.21 (1.45)	2.57 (1.14)	3.78 (0.66)	4.94 (1.71)
	Deepfake with original skin tone	5.68 (1.02)	4.67 1.38)	3.12 (1.22)	4.48 (0.61)	3.42 (1.87)
	Deepfake with changed skin tone	5.67 (1.04)	4.66 (1.38)	3.16 (1.15)	4.51 (0.61)	2.78 (1.61)
	3D avatar	4.65 (1.51)	6.16 (1.04)	3.94 (1.52)	4.92 (0.74)	4.13 (2.13)
Body of test-taker	No obfuscation	5.56 (1.04)	-	2.46 (1.08)	3.01 (0.45)	-
	Blurring	4.99 (1.13)	4.27 (1.21)	3.61 (1.33)	4.29 (0.67)	4.49 (1.89)
	Deepfake	5.22 (1.08)	5.09 (1.51)	4.31 (1.72)	4.88 (0.87)	3.97 (2.02)
Background of test-taker	No obfuscation	5.23 (1.19)	-	3.61 (0.66)	3.28 (0.48)	-
	Blurring	4.21 (1.25)	3.59 (1.28)	3.51 (1.39)	3.76 (0.72)	5.12 (1.73)
	Deepfake	3.56 (1.55)	6.06 (1.33)	5.45 (1.55)	5.02 (0.75)	5.23 (1.78)
People in background	No obfuscation	5.96 (0.94)	-	2.69 (1.42)	3.22 (0.53)	-
	Blurring	4.58 (1.37)	4.62 (1.64)	3.41 (1.55)	4.21 (0.81)	4.88 (1.91)
	Silhouette	3.85 (1.47)	5.97 (1.42)	4.49 (1.87)	4.77 (0.82)	4.49 (2.02)
	3D avatar	4.95 (1.28)	5.07 (1.63)	3.67 (1.65)	4.57 (0.84)	3.62 (2.03)

varied across both ROIs and methods. For instance, replacing a test-taker's face with a 3D avatar might be perceived as better for perceived privacy and fairness than using a realistic deepfake face, but it could compromise more cheating information. To statistically validate the findings from the figure, we conducted 12 mixed-effects models for three dependent variables across four ROIs, followed by post hoc pairwise comparisons[1]. Overall, the variability in dependent variables was relatively low *(<1.42)* across all ROIs, suggesting that the primary source of variability was the impact of obfuscation methods rather than individual differences. Below, we discuss key findings for the foreground (test-takers' face and body) and background (background and other people appearing in it) regions. Detailed statistics of pairwise comparisons can be found in Table 4.

Obfuscating foreground areas: Using advanced methods such as 3D avatars and deepfakes on the face can be perceived to provide better privacy *(all MMD>0.45, p<.001)* and fairness *(all MMD>0.54, p<.001)* than applying conventional blurring. However, although 3D avatars could offer better privacy *(all MMD>1.48, p<.001)* and fairness *(all MMD>0.77, p<.001)* compared to deepfakes, they might hide more cheating information *(all MMD>1.01, p<.001)* than the latter. Interestingly, a deepfake with a changed skin tone compared to the original skin tone may not significantly impact the dependent variables *(all MMD<0.06, p>.45)*. On the other hand, when obfuscating body parts, using deepfake to replace it with a covered outfit may be better on all dependent variables than merely blurring them *(all MMD>0.21, p<.001)*.

Obfuscating background areas: Of the two expert recommended methods (blurring and deepfake) for obscuring the static

[1]Note that the significance level (α) has been adjusted to .001 (i.e., $\approx$ .05/55) according to the Bonferroni correction, as a total of 55 pairwise comparisons were performed for our analysis; please refer to Table 4 for the detailed results of the pairwise comparisons

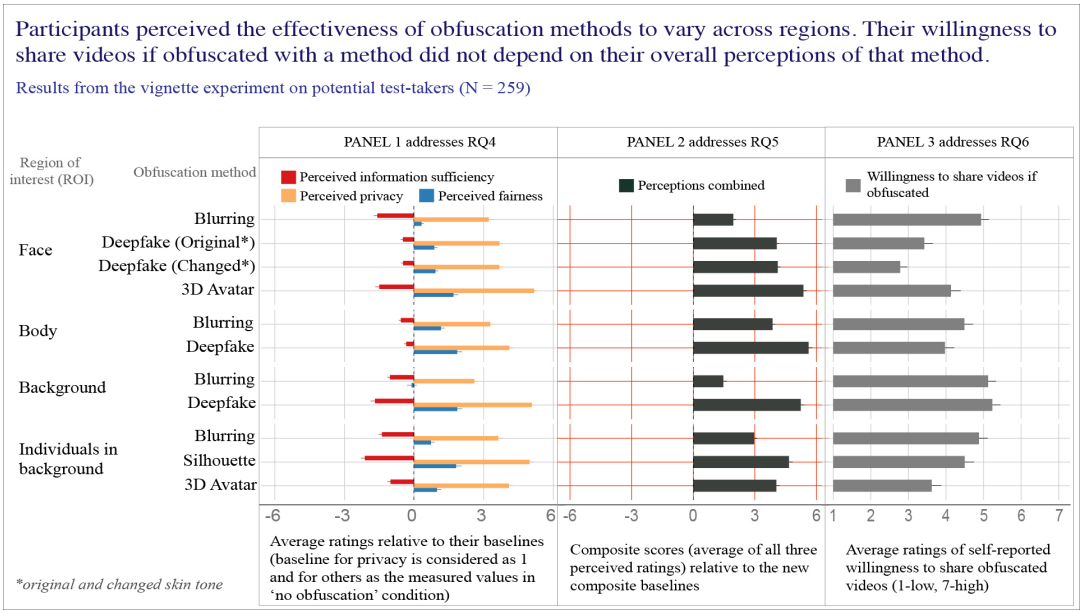

Figure 4: Three panels address three research questions respectively: the impact of region-specific obfuscation methods on *(RQ4)* perceptions of information sufficiency, privacy and fairness; *(RQ5)* the combined perception; and *(RQ6)* willingness to share videos if obfuscated. PANEL 1 and PANEL 2 plot bars relative to the respective baseline values

Table 4: Pairwise comparison of the effects of different ROI-specific obfuscation methods, presented as differences in marginal means (MMD) along with 95% confidence intervals and p-values for various dependent variables. Deepfake (O) and (C) represent deepfaking with original skin tone and changed skin tone respectively; ^{NS} denotes non-significant values at significance level=.05 and $^{NS+}$ denotes non-significant values at significance level=.001 after Bonferroni correction

Region of interest	Comparing effects between methods	Perceived information sufficiency	Perceived privacy protection	Perceived fairness	Composite scores	Willingness to share videos
Face of test-taker	Deepfake (O) - Blurring	1.09 [0.93, 1.26], p<.001	0.47 [0.28, 0.65], p<.001	0.55 [0.42, 0.68], p<.001	0.71 [0.62, 0.79], p<.001	-1.51 [-1.76, -1.27], p<.001
	Deepfake (C) - Blurring	1.11 [0.95, 1.27], p<.001	0.46 [0.25, 0.66], p<.001	0.59 [0.46, 0.73], p<.001	0.72 [0.63, 0.82], p<.001	-2.15 [-2.41, -1.91], p<.001
	3D avatar - Blurring	0.08 [-0.11, 0.26], p=.39^{NS}	1.96 [1.78, 2.14], p<.001	1.38 [1.18, 1.57], p<.001	1.14 [1.04, 1.24], p<.001	-0.81 [-1.05, -0.55], p<.001
	Deepfake (C) - Deepfake (O)	0.01 [-0.04, 0.06], p=.59^{NS}	-0.01 [-0.12, 0.11], p=.84^{NS}	0.05 [-0.07, 0.17], p=.46^{NS}	0.02 [-0.04, 0.08], p=.59^{NS}	-0.64 [-0.83, -0.45], p<.001
	3D avatar - Deepfake (O)	-1.02 [-1.16, -0.87], p<.001	1.49 [1.34, 1.63], p<.001	0.83 [0.63, 1.02], p<.001	0.43 [0.34, 0.53], p<.001	0.71 [0.46, 0.96], p<.001
	3D avatar - Deepfake (C)	-1.03 [-1.17, -0.89], p<.001	1.51 [1.35, 1.65], p<.001	0.78 [0.61, 0.96], p<.001	0.42 [0.33, 0.51], p<.001	1.35 [1.12, 1.58], p<.001
Body of test-taker	Deepfake - Blurring	0.22 [0.12, 0.33], p<.001	0.83 [0.63, 1.03], p<.001	0.69 [0.52, 0.87], p<.001	0.58 [0.49, 0.68], p<.001	-0.52 [-0.76, -0.28], p<.001
Background of test-taker	Deepfake - Blurring	-0.65 [-0.81, -0.49], p<.001	2.47 [2.24, 2.71], p<.001	1.94 [1.73, 2.16], p<.001	1.26 [1.14, 1.37], p<.001	0.11 [-0.11, 0.32], p=.31^{NS}
People in background	Silhouette - Blurring	-0.73 [-0.85, -0.62], p<.001	1.35 [1.18, 1.51], p<.001	1.08 [0.91, 1.26], p<.001	0.57 [0.48, 0.65], p<.001	-0.38 [-0.61, -0.16], p<.001
	3D avatar - Blurring	0.37 [0.23, 0.51], p<.001	0.46 [0.25, 0.66], p<.001	0.26 [0.08, 0.44], p<.01^{NS+}	0.37 [0.26, 0.46], p<.001	-1.25 [-1.51, -0.99], p<.001
	3D avatar - Silhouette	1.11 [0.95, 1.25], p<.001	-0.89 [-1.07, -0.72], p<.001	-0.83 [-1.02, -0.64], p<.001	-0.21 [-0.31, -0.11], p<.001	-0.88 [-1.09, -0.66], p<.001

background, using deepfake to replace it with an image can be perceived as a better method for both privacy *(MMD=2.47, p<.001)* and fairness *(MMD=1.94, p<.001)* - than just blurring it. But the former may hide more cheating information than the latter *(MMD=0.65, p<.001)*. Also, if people show up in the background, replacing them with a silhouette-like figure can be found to be most effective for both privacy *(all MMD>0.88, p<.001)* and fairness *(all MMD>0.82, p<.001)*, better than using either a full-body 3D avatar or blurring. Yet, using the silhouette could compromise on retaining cheating information more than either a full-body 3D avatar or blurring *(all MMD>0.72, p<.001)*. Based on the analysis conducted so far, no single obfuscation method emerged as the best choice across all three dependent variables across ROIs. Hence, we will proceed to examine the effects of obfuscation methods on the overall perception.

5.2.2 Effects of Region-specific Obfuscation Methods on the Perceptions Combined. To address RQ5, which examines the impact of obfuscation methods on participants' overall perceptions, we created composite scores by averaging the ratings of all three dependent variables. Considering the variables were significantly correlated to each other (*most have r>0.15, p<.001*; see Appendix Table 6), composite scores were only used to measure the overall perception without encompassing a broader concept. PANEL 2 in Figure 4 presents these composite scores compared to the new

composite baselines. We ran four mixed-effects models followed by post hoc pairwise comparisons. In all models, variability was low (*<0.36*) across ROIs, indicating that obfuscation methods had a greater impact than individual differences.

Obfuscating foreground areas: When obscuring test-takers' face and body, the overall perceptions of the obfuscation method's effectiveness can be influenced more by perceived privacy *(all r>0.67, p<.001)* and fairness *(all r>0.71, p<.001)*, than by the amount of cheating information being suppressed *(all r>0.16, p<.001)*. Using a 3D avatar for the face can perform significantly better overall than both deepfake and blurring *(all MMD>0.41, p<.001)*. Changing the face color with deepfake may not significantly alter perceptions *(MMD=0.02, p=.59)*. For body obfuscation, using deepfake to change the outfit can significantly provide better overall perception than just blurring it *(MMD=0.58, p<.001)*.

Obfuscating background areas: Similar to foreground areas, when we obscure background elements, test-takers may assess the methods' effectiveness more based on their perceptions of privacy *(all r>0.64, p<.001)* and fairness *(all r>0.76, p<.001)* rather than the extent of cheating information suppression *(all r>0.03, p<.001)*. Using deepfake to replace a static background can produce better results compared to merely blurring it *(MMD=1.26, p<.001)*. Additionally, replacing individuals in the background, if present, with a silhouette-like figure can prove more effective than employing either a 3D avatar or blurring techniques *(all MMD>0.21, p<.001)*. Having identified which obfuscation methods work best for specific ROIs based on the overall perceptions, we'll next explore if participants' UX aligns with those perceptions.

5.2.3 Effects of Region-specific Obfuscation on Willingness to Share Obfuscated Video. PANEL 3 in Figure 4 addresses RQ6 by showing average willingness ratings for each region-specific obfuscation method. We used four mixed-effects models for four ROIs, followed by post hoc analyses. The variability was moderately low across all ROIs *(<2.09)*, indicating that obfuscation methods had a greater impact on the willingness ratings than participants' individual differences. Key findings are reported below.

In the face region, blurring was significantly more preferred than either deepfakes or 3D avatars *(all MMD>0.8, p<.001)*. Interestingly, participants were less willing to share their videos when deepfaked with skin color change compared to without skin color change *(MMD=0.64, p<.001)*. A similar preference for blurring was found for body and background individuals, except in the background where both blurring and deepfake had high ratings with no significant difference *(MMD=0.11, p=.31)*. Contrary to the findings in Section 5.2.2, where participants' overall perceptions favored advanced obfuscation methods (e.g., 3D avatar, deepfake), they predominantly opted for blurring in all ROIs for sharing videos with reviewers. Since the overall perceptions didn't match their sharing preferences, we checked for correlations and found either no relationship for foreground areas (face and body) *(all r<0.04, p>.45)* or low correlation *(all r<0.2, p<.001)* for background areas (still background and people). This will be discussed further in the discussion section.

5.2.4 Qualitative Analysis of Scope for Discrimination Post Obfuscation. Our study design accounted for potential discriminatory factors (skin tone, ethnicity and gender) based on expert recommendations while creating test stimuli. Hence, we analyzed the open-ended questions using an inductive coding process in MAXQDA (v. 2024) to understand participants' perceptions of obfuscation addressing these factors. The first author generated codes during the inductive coding process, given the short and straightforward nature of the data (codes are provided in Appendix Table 8).

We have already observed in Section 5.2.1 and 5.2.2 that the alteration of participants' skin tone while using the deepfake method didn't show significant effects. However, more participants (around one-third) expressed concerns in open-ended responses regarding the alteration of a fair-skinned test-taker to a darker skin color, compared to one-fifth in the reverse scenario. They cited that altered skin tone could bias reviewers' perceptions of assumed ethnicity. Next, we didn't statistically compare the obfuscation method's effectiveness across genders due to simultaneous alterations in the test stimuli beyond gender, e.g., body, background. However, open-ended responses highlighted persistent cues implying the assumed gender of test-takers despite obfuscation in certain ROIs. Concerns often centered around unchanged length and textures of hair in the deepfaked face; gender-specific outfits chosen for body replacement using deepfake; and differences in body builds even after replacing the background individuals with a silhouette. Using 3D avatars for replacing test-takers' faces or background individuals' full bodies also raised concerns regarding the selected avatar color and gender alignment, particularly when background individuals were replaced with avatars of the same gender.

Fairness concerns were also reported beyond biases related to skin tone, ethnicity and gender. For instance, inferring test-takers' socio-economic status from unconventional test-taking places (e.g., a warehouse used in our study design) or perceiving the presence of background individuals during test-taking as unprofessional could potentially lead to discrimination. Other concerns arose like potential distraction caused by visually appealing 3D avatars for background individuals. Furthermore, suspicions of cheating may arise from complete background replacement with a picture or substituting background individuals with a silhouette, potentially undermining the fairness of obfuscation measures.

6 Discussion

This paper addresses key challenges in remote proctoring that may affect test-taking experience. By exploring promising obfuscation methods that can hide privacy-sensitive details in video recordings, the study aims to improve test-takers' experience, for example, their perceptions on privacy protection. It further explores whether these obfuscation methods can eliminate potential discriminatory attributes in the videos (e.g., test-takers' ethnicity, gender) [57], which may unfairly influence reviewers' judgments and lead to unjust accusations of cheating. Finally, these methods must not compromise the core purpose of remote proctoring, which is to provide reviewers with videos containing enough information to detect cheating. Test-takers' perception of this information sufficiency can influence their test-taking experience and their likelihood of cheating, assuming they believe that cheating actions can be identified.

In this study, we interviewed experts from e-assessment, computer vision and usable privacy fields to identify the most promising obfuscation methods for different video regions, such as test-takers' face, body, background and people in the background. We then examined how these methods affect non-expert potential test-takers' perceptions of privacy protection, fairness and information sufficiency, as well as their willingness to adopt them for remote proctoring.

We explored obfuscation methods with distinct working principles, potentially evoking varied user perceptions [29]. For instance, blurring gradually fades content like faces; deepfake merges content characteristics with a reference to create a new realistic output; 3D avatar overlays a reference to fully conceal the content; and silhouette replacement blacks out content based on its contour, potentially revealing the original content due to the visible outline.

Acknowledging the computational limitations of implementing promising methods like deepfake in videos, we simplified our user testing by using simulated images instead. Participants were instructed to imagine these images as visual scenes in video recordings. While static images may not fully capture the dynamic nature of video content, they can still encapsulate crucial elements influencing user perceptions [8, 41]. As these promising obfuscation techniques mature, they may become more viable for integration into video editing workflows in the future. Below, we provide insights on obfuscations, contributing to the discourse surrounding standardizing video recording in remote proctoring.

6.1 Obfuscating Foreground Areas of Remote Proctored Videos

Previous studies [5, 29] found that obscuring both the face and body of test-takers provides greater privacy protection than just obscuring the face. However, in our study, we did not uniformly obfuscate the full bodies of test-takers. Instead, we evaluated different obfuscation methods separately for the face and body because different regions of a video may offer different privacy-utility trade-offs [11, 39]. For example, some regions may have high privacy significance, while others may be most informative for cheating detection (utility). Additionally, our study considers fairness in cheating detection as another important factor in this trade-off, alongside privacy and cheating detection.

The test-taker's face is highly identifiable, raising significant privacy and fairness concerns. Obfuscating the face is challenging because facial expressions, head pose and eye gaze are crucial for detecting cheating behavior. While deepfake with realistic face replacement can commonly be used for this purpose [48], our evaluation by potential test-takers (in Section 5.2.2) indicated that a cartoonized 3D avatar replacement could be more effective, based on the overall perceptions of privacy, fairness and information sufficiency. This relative effectiveness of 3D avatars reflected in higher privacy and fairness ratings (see Figure 4), aligns with a prior study [29] indicating similar privacy benefits with avatarized face representations. However, it's unclear if this effect was due to the obfuscation methods themselves or subtle differences between their designs. For example, during stimuli design, the gender-neutral cartoon face used in 3D avatars contrasts with deepfake methods that retain the test-takers' original facial characteristics (e.g., hair), potentially prompting assumptions about their gender or ethnicity and raising fairness concerns. Additionally, the adjustable size of 3D avatars can hide further cues, such as neck color, thus mitigating assumptions about the test-taker's ethnicity.

Body obfuscation is also important because the body can reveal sensitive cues such as attire type or tattoos [6, 23]. The best way to conceal these cues might involve replacing the entire body region with a realistically covered outfit. Deepfake is a promising candidate for this, as it preserves motion information [39], such as frequent hand or shoulder movements, crucial for cheating detection. Collectively, our findings imply that achieving an optimal level of obfuscation in remote proctoring videos may entail applying distinct obfuscation techniques for the face and body regions, while simultaneously ensuring the preservation of critical motion information essential for detecting cheating behaviors.

6.2 Obfuscating Background Areas of Remote Proctored Videos

Obfuscation of background areas poses a unique challenge as it involves concealing sensitive visual cues such as living conditions, specific objects or the presence of other individuals (e.g., family members) [46] without losing relevant information that may assist test-takers. Background obfuscation requires a distinct strategy for static backgrounds and individuals appearing in the background. Unlike the separate obfuscation for the test-takers' face and body, as discussed in Section 6.1, the lower frequency of individuals appearing in the background may allow for a full-body obfuscation.

A promising approach could involve replacing the static background with a generic picture for all test-takers using deepfake, similar to video conferencing tools, while excluding any individuals detected in the background. This also mitigates potential discrimination against test-takers based on factors like social status or lack of professional background, especially when tests are taken in unconventional places, as discussed in Section 5.2.4. Next, adhering strictly to remote testing guidelines for an ideal test-taking environment without visible individuals may be impractical for test-takers with diverse living situations. However, if interactions between test-takers and those individuals are strictly prohibited during test-taking, a silhouette-like figure replacing their entire body, while preserving motion and contour, can still be informative for cheating detection while ensuring privacy protection. However, caution is needed, as the silhouette's contour may lead to assumptions about their gender, potentially undermining fairness expectations.

6.3 Visual Obfuscation - a Preferred Solution for Remote Proctoring?

State-of-the-art obfuscation methods (e.g., deepfake, 3D avatar) appear promising in offering test-takers adequate levels of perceived privacy protection, perceived fairness and perceived information sufficiency. However, non-expert participants expressed hesitancy in sharing their videos with unknown reviewers, particularly when obfuscating their faces or bodies using these methods. In contrast, blurring emerged as the preferred solution for both regions compared to those advanced methods. This resonates with the findings from a prior study [29] on applying obfuscation methods to social

media photos, where solutions offering lower privacy but higher information sufficiency were favored. Our study also considered fairness aspects alongside privacy and cheating detection in the context-dependent evaluation of the privacy-utility trade-off [50]. Based on the willingness ratings, it appears that the perceived effectiveness of those advanced techniques (i.e., deepfake, 3D avatar) in retaining crucial cheating information may be lacking, or there might be a lack of trust [10, 43] in their practical implementation. Test-takers may also have concerns about potential video glitches, issues related to avatar design, or the selection of gender and skin tone for face or body replacement, that can affect obfuscation effectiveness. Blurring may also be preferred due to its familiarity and cost-effectiveness. Hence, addressing these concerns could bolster test-takers' trust in more advanced obfuscation methods and encourage them to share obfuscated videos more willingly. Taken together, our study offers insights into the expected outcomes when implementing obfuscation techniques on video recordings, paving the way for further research, as outlined in Section 8.

6.4 Safeguarding Privacy of the Unprocessed Video Files

Figure 1 outlines the expert-recommended proctoring pipeline, where different region-specific obfuscation methods can be applied to the copies of the video recordings for review. It is also important to preserve the privacy of the unprocessed videos. In practice, video recordings are typically stored for several weeks or months before deletion, as mandated by test-organizing institutes [6]. If long-term retention of unprocessed videos is intended, institutes may also consider a temporal redaction approach [37], which can obfuscate different video regions after a specific period of time. This, however, needs further research into the appropriateness of effective redaction strategies in this context.

7 Actionable Guidelines for Researchers and Practitioners

Our study suggests insights with following guidelines for balancing test-takers' perceptions of privacy and fair cheating detection in remote proctoring setups, particularly if obfuscation solutions are applied during post-test video processing.

(1) **Effective obfuscation methods for distinct regions**:
   (a) Foreground region: Replacing a test-taker's face with a sufficiently large uniform 3D avatar face can effectively hide facial features, including the neck, while retaining real-time expressions like eye gaze, mouth and head movements, crucial for successful cheating detection. Next, a standardized gender-neutral professional outfit, such as a formal suit, can effectively replace the body region, preserving the movement of hands and shoulders.
   (b) Background region: Choosing a uniform generic background image for replacement can effectively conceal test-taking environment and bias-inducing elements, but test-organizing institutes should verify for presence of cameras behind test-takers in unprocessed videos after reviewers assess the obfuscated videos, crucial for cheating detection. Individuals in the background can be replaced with a mono-colored 2D silhouette-like figure by accurately measuring their body contour and retaining real-time body movements.

(2) **Practical remote proctoring pipeline**: A hybrid remote proctoring approach combines algorithmic-based cheating detection with manual verification. Recorded videos should be securely stored with a trusted entity; relevant obfuscation methods can be applied to a copy of the recorded videos and shared with external reviewers hired by test-organizing institutes (refer to Figure 1). Only the institutes would have access to the unprocessed videos for a final review based on reports from reviewers, to resolve potential disputes if test-takers challenge cheating allegations.

8 Limitation and Future Scope

While this study highlights the potential of region-specific obfuscation methods to improve the test-taking experiences, future research could expand upon this work to fully understand and realize their benefits in remote proctoring.

First, while our survey relied on static images to illustrate obfuscation methods, future studies could apply these methods directly to video recordings of test-takers to provide participants with a more realistic portrayal. However, this approach would require addressing various technical, logistical, legal and ethical challenges. Furthermore, well-constructed guidelines aimed at ensuring fairness could also be developed to address any additional fairness concerns that may arise, particularly regarding modified regions in the videos post-obfuscation. *Second*, exploring the perspectives of proctoring managers could offer valuable insights into the practical challenges and feasibility of implementing obfuscation techniques in real-world scenarios. *Third*, we constructed scales (Table 5) for test-takers' perception of privacy, cheating information sufficiency and fairness using ad-hoc items. The development and validation of standardized scales for these dimensions could greatly benefit researchers in this area. *Finally*, while our survey sampled a diverse group of participants as potential test-takers, future studies could target specific populations such as students or professionals to explore potential differences in their perceptions and experiences with remote proctoring.

9 Conclusion

This study marks an initial step toward enhancing test-takers' experiences in remote proctoring by addressing key concerns surrounding the review of proctored videos: protecting their privacy and accurate and fair cheating detection. Through the selective obfuscation of privacy-sensitive visual data in various video regions (test-taker's face, body, background and individuals in the background) using expert-recommended (N=9) obfuscation methods, we evaluated their impact on potential test-takers' (N=259) overall perceptions of those three dimensions as well as their willingness to share obfuscated videos. Our findings underscore the significant impact of region-specific obfuscation methods on participant experiences, suggesting that optimal outcomes may be achieved by tailoring obfuscation methods across regions (e.g., 3D avatar on the face, deepfake on the body, silhouette on background individuals). We provide guidance for researchers and practitioners to assess

the cost-effectiveness of testing these methods with real proctored videos before practical implementation, while also advocating for further research in this pertinent and evolving field.

Acknowledgments

This research is the result of the project 'Secure and Verifiable Electronic Testing and Assessment Systems' (INTER/ANR/20 /14926102/SEVERITAS) funded by the Luxembourg National Research Fund (FNR) and the French National Research Agency (ANR).

References

[1] David G Balash, Dongkun Kim, Darika Shaibekova, Rahel A Fainchtein, Micah Sherr, and Adam J Aviv. 2021. Examining the examiners: Students' privacy and security perceptions of online proctoring services. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*. 633–652.

[2] Ben Burgess, Avi Ginsberg, Edward W Felten, and Shaanan Cohney. 2022. Watching the watchers: bias and vulnerability in remote proctoring software. In *31st USENIX Security Symposium (USENIX Security 22)*. 571–588.

[3] Chris Burt. 2019. Sensetime partners with China Tower for massive biometric video surveillance network. *Biometric Update* (30 Sep 2019). https://www.biometricupdate.com/201909/sensetime-partners-with-china-tower-for-massive-biometric-video-surveillance-network

[4] Wm Camron Casper, Bryan D Edwards, J Craig Wallace, Ronald S Landis, and Dustin A Fife. 2020. Selecting response anchors with equal intervals for summated rating scales. *Journal of Applied Psychology* 105, 4 (2020), 390.

[5] Datong Chen, Yi Chang, Rong Yan, and Jie Yang. 2009. Protecting personal identification in video. *Protecting Privacy in Video Surveillance* (2009), 115–128.

[6] Simon Coghlan, Tim Miller, and Jeannie Paterson. 2021. Good proctor or "big brother"? Ethics of online exam supervision technologies. *Philosophy & Technology* 34, 4 (2021), 1581–1606.

[7] Rianne Conijn, Ad Kleingeld, Uwe Matzat, and Chris Snijders. 2022. The fear of big brother: The potential negative side-effects of proctored exams. *Journal of Computer Assisted Learning* (2022).

[8] Claire Anne Conway, Benedict Christopher Jones, Lisa M DeBruine, and Anthony C Little. 2008. Evidence for adaptive design in human gaze preference. *Proceedings of the Royal Society B: Biological Sciences* 275, 1630 (2008), 63–69.

[9] Mohammad Dadashzadeh. 2021. The Online Examination Dilemma: To Proctor or Not to Proctor?. *Journal of Instructional Pedagogies* 25 (2021).

[10] Catherine Dwyer, Starr Hiltz, and Katia Passerini. 2007. Trust and privacy concern within social networking sites: A comparison of Facebook and MySpace. *AMCIS 2007 proceedings* (2007), 339.

[11] Adám Erdélyi, Thomas Winkler, and Bernhard Rinner. 2018. Privacy protection vs. utility in visual data: An objective evaluation framework. *Multimedia tools and applications* 77 (2018), 2285–2312.

[12] Thomas B Fitzpatrick. 1988. The validity and practicality of sun-reactive skin types I through VI. *Archives of dermatology* 124, 6 (1988), 869–871.

[13] Tina L Freiburger, Danielle M Romain, Blake M Randol, and Catherine D Marcum. 2017. Cheating behaviors among undergraduate college students: Results from a factorial survey. *Journal of Criminal Justice Education* 28, 2 (2017), 222–247.

[14] Yun Fu, Renxiang Li, Thomas S Huang, and Mike Danielsen. 2008. Real-time multimodal human–avatar interaction. *IEEE Transactions on Circuits and Systems for Video Technology* 18, 4 (2008), 467–477.

[15] Kelly Grindstaff and Michael Mascarenhas. 2019. "No One Wants to Believe It": Manifestations of White Privilege in a STEM-Focused College. *Multicultural Perspectives* 21, 2 (2019), 102–111.

[16] Marc Hassenzahl and Noam Tractinsky. 2006. User experience-a research agenda. *Behaviour & information technology* 25, 2 (2006), 91–97.

[17] Darragh Higgins, Rebecca Fribourg, and Rachel McDonnell. 2021. Remotely perceived: Investigating the influence of valence on self-perception and social experience for dyadic video-conferencing with personalized avatars. *Frontiers in Virtual Reality* 2 (2021), 668499.

[18] Steven Hill, Zhimin Zhou, Lawrence K Saul, and Hovav Shacham. 2016. On the (In) effectiveness of Mosaicing and Blurring as Tools for Document Redaction. *Proc. Priv. Enhancing Technol.* 2016, 4 (2016), 403–417.

[19] Mohammed Juned Hussein, Javed Yusuf, Arpana Sandhya Deb, Letila Fong, and Som Naidu. 2020. An evaluation of online proctoring tools. *Open Praxis* 12, 4 (2020), 509–525.

[20] Tabitha L James, Linda Wallace, Merrill Warkentin, Byung Cho Kim, and Stéphane E Collignon. 2017. Exposing others' information on online social networks (OSNs): Perceived shared risk, its determinants, and its influence on OSN privacy control use. *Information & Management* 54, 7 (2017), 851–865.

[21] Ana Javornik, Ben Marder, Jennifer Brannon Barhorst, Graeme McLean, Yvonne Rogers, Paul Marshall, and Luk Warlop. 2022. 'What lies behind the filter?'Uncovering the motivations for using augmented reality (AR) face filters on social media and their effect on well-being. *Computers in Human Behavior* 128 (2022), 107126.

[22] Dima Kagan, Galit Fuhrmann Alpert, and Michael Fire. 2023. Zooming Into Video Conferencing Privacy. *IEEE Transactions on Computational Social Systems* (2023).

[23] Faten F Kharbat and Ajayeb S Abu Daabes. 2021. E-proctored exams during the COVID-19 pandemic: A close understanding. *Education and Information Technologies* 26, 6 (2021), 6589–6605.

[24] Yeolib Kim and Robert A Peterson. 2017. A Meta-analysis of Online Trust Relationships in E-commerce. *Journal of interactive marketing* 38, 1 (2017), 44–54.

[25] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. 2019. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8878–8887.

[26] Karen Lander, Vicki Bruce, and Harry Hill. 2001. Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition* 15, 1 (2001), 101–116.

[27] Haotian Li, Min Xu, Yong Wang, Huan Wei, and Huamin Qu. 2021. A visual analytics approach to facilitate the proctoring of online exams. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–17.

[28] Yifang Li and Kelly Caine. 2022. Obfuscation Remedies Harms Arising from Content Flagging of Photos. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–25.

[29] Yifang Li, Nishant Vishwamitra, Bart P Knijnenburg, Hongxin Hu, and Kelly Caine. 2017. Effectiveness and users' experience of obfuscation as a privacy-enhancing technology for sharing photos. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–24.

[30] Torrin M Liddell and John K Kruschke. 2018. Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology* 79 (2018), 328–348.

[31] Morgan Meaker. 2023. This Student Is Taking On 'Biased' Exam Software. *Wired* (5 Apr 2023). https://www.wired.co.uk/article/student-exam-software-bias-proctorio

[32] Suvadeep Mukherjee, Björn Rohles, Verena Distler, Gabriele Lenzini, and Vincent Koenig. 2023. The effects of privacy-non-invasive interventions on cheating prevention and user experience in unproctored online assessments: An empirical study. *Computers & Education* 207 (2023), 104925.

[33] Geoff Norman. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education* 15 (2010), 625–632.

[34] Paraskevi Nousi, Sotirios Papadopoulos, Anastasios Tefas, and Ioannis Pitas. 2020. Deep autoencoders for attribute preserving face de-identification. *Signal Processing: Image Communication* 81 (2020), 115699.

[35] Eyal Peer, David Rothschild, Andrew Gordon, Zak Evernden, and Ekaterina Damer. 2022. Data quality of platforms and panels for online behavioral research. *Behavior Research Methods* (2022), 1.

[36] Ian S Penton-Voak and Helen Y Chang. 2008. Attractiveness judgements of individuals vary across emotional expression and movement conditions. *Journal of Evolutionary Psychology* 6, 2 (2008), 89–100.

[37] Sabid Bin Habib Pias, Imtiaz Ahmad, Taslima Akter, Apu Kapadia, and Adam J Lee. 2022. Decaying Photos for Enhanced Privacy: User Perceptions Towards Temporal Redactions and'Trusted'Platforms. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–30.

[38] Tejaswi Potluri and Venkata Krishna Kishore K. 2023. An automated online proctoring system using attentive-net to assess student mischievous behavior. *Multimedia Tools and Applications* (2023), 1–30.

[39] Siddharth Ravi, Pau Climent-Pérez, and Francisco Florez-Revuelta. 2023. A review on visual privacy preservation techniques for active and assisted living. *Multimedia Tools and Applications* (2023), 1–41.

[40] Jessica L Roberts. 2014. Protecting privacy to prevent discrimination. *Wm. & Mary L. Rev.* 56 (2014), 2097.

[41] S Craig Roberts, Tamsin K Saxton, Alice K Murray, Robert P Burriss, Hannah M Rowland, and Anthony C Little. 2009. Static and dynamic facial images cue similar attractiveness judgements. *Ethology* 115, 6 (2009), 588–595.

[42] Somdev Sangwan. 2023. One-click face swap. https://github.com/s0md3v/roop.

[43] Simeon Schudy and Verena Utikal. 2017. 'You must not know about me'—On the willingness to share personal data. *Journal of Economic Behavior & Organization* 141 (2017), 1–13.

[44] Valentin Schwind, Katrin Wolf, Niels Henze, and Oliver Korn. 2015. Determining the characteristics of preferred virtual faces using an avatar generator. In *proceedings of the 2015 annual symposium on computer-human interaction in play*. 221–230.

[45] Neil Selwyn, Chris O'Neill, Gavin Smith, Mark Andrejevic, and Xin Gu. 2023. A necessary evil? The rise of online exam proctoring in Australian universities. *Media International Australia* 186, 1 (2023), 149–164.

[46] Arnout Terpstra, Alwin De Rooij, and Alexander Schouten. 2023. Online Proctoring: Privacy Invasion or Study Alleviation? Discovering Acceptability Using

Contextual Integrity. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.
[47] William Thong, Przemyslaw Joniak, and Alice Xiang. 2023. Beyond skin tone: A multidimensional measure of apparent skin color. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4903–4913.
[48] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. 2020. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion* 64 (2020), 131–148.
[49] Ehsan Ul Haque, Mohammad Maifi Hasan Khan, and Md Abdullah Al Fahim. 2023. The Nuanced Nature of Trust and Privacy Control Adoption in the Context of Google. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–23.
[50] André Calero Valdez and Martina Ziefle. 2019. The users' perspective on the privacy-utility trade-offs in health recommender systems. *International Journal of Human-Computer Studies* 121 (2019), 108–121.
[51] Xinrui Wang and Jinze Yu. 2020. Learning to cartoonize using white-box cartoon representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8090–8099.
[52] Sera Whitelaw, Mamas A Mamas, Eric Topol, and Harriette GC Van Spall. 2020. Applications of digital technology in COVID-19 pandemic planning and response. *The Lancet Digital Health* 2, 8 (2020), e435–e440.
[53] Skye Witley. 2023. Virtual Exam Case Primes Privacy Fight on College Room Scans. *Bloomberg Law* (25 Jan 2023). https://news.bloomberglaw.com/privacy-and-data-security/virtual-exam-case-primes-privacy-fight-over-college-room-scans
[54] Rongbin Yang and Santoso Wibowo. 2022. User trust in artificial intelligence: A comprehensive conceptual framework. *Electronic Markets* 32, 4 (2022), 2053–2077.
[55] Shiyu Yang, Dominique Brossard, Dietram A Scheufele, and Michael A Xenos. 2022. The science of YouTube: What factors influence user engagement with online science videos? *Plos one* 17, 5 (2022), e0267697.
[56] Waheeb Yaqub, Manoranjan Mohanty, and Basem Suleiman. 2022. Privacy-Preserving Online Proctoring using Image-Hashing Anomaly Detection. In *2022 International Wireless Communications and Mobile Computing (IWCMC)*. IEEE, 1113–1118.
[57] Deborah R Yoder-Himes, Alina Asif, Kaelin Kinney, Tiffany J Brandt, Rhiannon E Cecil, Paul R Himes, Cara Cashon, Rachel MP Hopp, and Edna Ross. 2022. Racial, skin tone, and sex disparities in automated proctoring software. In *Frontiers in Education*, Vol. 7. Frontiers, 881449.
[58] Lin Yuan, Linguo Liu, Xiao Pu, Zhao Li, Hongbo Li, and Xinbo Gao. 2022. PRO-Face: A Generic Framework for Privacy-preserving Recognizable Obfuscation of Face Images. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1661–1669.
[59] Yan Zhuang. 2018. The performance cost of software obfuscation for Android applications. *Computers & Security* 73 (2018), 57–72.

A Materials used in Non-expert Evaluation

A.1 Supplementary Files

The files containing the complete survey and the dataset used for analysis can be accessible at: https://osf.io/3prv8/?view_only=40c7e78d248342e8b770d7098e923360

A.2 Pilot Studies for Non-expert Evaluation

While modifying the created stimuli by blurring different regions of interest (ROI) e.g., face, body, background and people in the background, a brief pilot test involving six participants was conducted. They adjusted the radius of the blurred ROI using a slider (ranging from 1-100), aiming for a balance between effective ROI concealment and visual information sufficiency for cheating detection. The median value of the collected blurred radius for each ROI was applied to the modified stimuli. We also conducted another pilot study of the entire survey with 10 participants recruited from Prolific, using a feedback box to identify any issues with the questions or image quality. Since no changes were made to the survey afterward, the responses from this pilot test were included in the data analysis.

A.3 Details of Items used in Survey

Table 5: Measurements during vignette experiment with non-experts

Scale	Region of Interest (ROI)	Items used in the survey with 7-point Likert scale on agreement. The levels used are: *Strongly disagree, Disagree, Somewhat disagree, Neither disagree nor agree, Somewhat agree; Agree, Strongly agree*
Perceived information sufficiency	Face	1. Reviewers could detect when the test-taker carries out mouth movement, such as talking. 2. Reviewers could detect when the test-taker interacts with other people (in the room). 3. Reviewers could detect when the test-taker looks away from the screen.
	Body	1. Reviewers could detect test-taker's body movements suggesting unauthorized use of resources (e.g smartphone). 2. Reviewers could detect when the test-taker moves from their seat.
	Background	1. Reviewers could detect when the test-taker interacts with other people in their environment. 2. Reviewers could detect the presence of a camera in the background recording the test-taker's computer screen.
	People in background	1. Reviewers could detect when the person in the background talks. 2. Reviewers could detect when the test-taker interacts with the person in the background.
Perceived privacy concerns	Face	1. Reviewers could recognize the identity of the test-taker.
	Body	1. Reviewers could recognize the outfits of the test-taker. 2. Reviewers could recognize the body identity of the test-taker.
	Background	1. Reviewers could recognize the objects present in the background. 2. Reviewers could recognize the location of the test-taker.
	People in background	1. Reviewers could recognize the identity of the person in the background.
Perceived fairness concerns	All four ROIs	1. Biased reviewers could exhibit discrimination against the test-taker.
		(Open-ended question) What aspects of the test-taker's modified image might be used by a biased reviewer to discriminate against this test-taker?
Willingness to video sharing	All four ROIs	1. I am willing to share my modified video with the reviewers.

A.4 Correlations Table for Non-expert Evaluation

Table 6: Pairwise Correlations between dependent variables. Pearson's correlation coefficient. $^{+}$p<.1, *p<.05, **p<.01, ***p<.001

ROI	Dependent variables	(1)	(2)	(3)	(4)	(5)
Face	(1): Perceived privacy	1.00				
	(2): Perceived info. sufficiency	-0.29***	1.00			
	(3): Perceived fairness	0.31***	-0.15***	1.00		
	(4): Composite score	0.66***	0.31***	0.69***	1.00	
	(5): Willingness to video sharing	-0.02	-0.01	0.01	-0.01	1.00
Body	(1): Perceived privacy	1.00				
	(2): Perceived info. sufficiency	-0.25***	1.00			
	(3): Perceived fairness	0.24***	-0.06^{+}	1.00		
	(4): Composite score	0.61***	0.34***	0.79***	1.00	
	(5): Willingness to video sharing	-0.13**	0.18	0.05	0.03	1.00
Background	(1): Perceived privacy	1.00				
	(2): Perceived info. sufficiency	-0.44***	1.00			
	(3): Perceived fairness	0.64***	-0.34***	1.00		
	(4): Composite score	0.78***	0.01	0.84***	1.00	
	(5): Willingness to video sharing	0.07^{+}	-0.01	0.19***	0.16***	1.00
People in background	(1): Perceived privacy	1.00				
	(2): Perceived info. sufficiency	-0.45***	1.00			
	(3): Perceived fairness	0.37***	-0.26***	1.00		
	(4): Composite score	0.64***	0.09**	0.77***	1.00	
	(5): Willingness to video sharing	0.16***	-0.07^{+}	0.18***	0.19***	1.00

A.5 Regression Table from Non-expert Evaluation

Table 7: Effects of different region-specific obfuscation methods on the dependent variables. Since ratings without obfuscation were only collected for perceived information sufficiency and fairness, the mixed-effects model coefficients along with 95% CI for these variables are presented relative to the 'no obfuscation' condition as the baseline. For the remaining dependent variables, the coefficients along with 95% CI are shown relative to 'blurring' as the baseline. The baselines are shown in bold font

Region of interest	Obfuscation methods	Perceived information sufficiency	Perceived privacy protection	Perceived fairness	Composite scores	Willingness to share videos
Face of test-taker	Deepfake with~ original skin tone	-0.47*** [-0.62, -0.33]	0.47*** [0.28, 0.65]	0.86*** [0.71, 1.02]	0.71***[0.62, 0.79]	-1.51*** [-1.75, -1.26]
	Deepfake with~ changed skin tone	-0.46*** [-0.61, -0.31]	0.46*** [0.23, 0.66]	0.91*** [0.75, 1.07]	0.72***[0.63, 0.82]	-2.15*** [-2.41, -1.91]
	3D avatar	-1.49*** [-1.63, -1.34]	1.96*** [1.78, 2.14]	1.69*** [1.53, 1.85]	1.14***[1.04, 1.24]	-0.81*** [-1.05, -0.55]
	Blurring	-1.57*** [-1.71, -1.42]	**4.21 [4.03, 4.38]**	0.31*** [0.15, 0.47]	**3.78 [3.69, 3.86]**	**4.93 [4.72, 5.14]**
	No Obfuscation	**6.14 [6.03, 6.24]**	-	**2.25 [2.12, 2.39]**	-	-
Body of test-taker	Deepfake	-0.33*** [-0.43, -0.23]	0.83*** [0.63, 1.03]	1.85*** [1.66, 2.03]	0.58***[0.49, 0.68]	-0.52*** [-0.76, -0.28]
	Blurring	-0.56*** [-0.66, -0.45]	**4.27 [4.12, 4.42]**	1.15*** [0.64, 1.34]	**4.29 [4.21, 4.37]**	**4.49 [4.25, 4.72]**
	No Obfuscation	**5.55 [5.43, 5.68]**	-	**2.46 [2.33, 2.59]**	-	-
Background of test-taker	Deepfake	-1.68*** [-1.84, -1.52]	2.47*** [2.24, 2.71]	1.85*** [1.63, 2.06]	1.25***[1.14, 1.37]	0.11 [-0.09, 0.32]
	Blurring	-1.03*** [-1.19, -0.87]	**3.59 [3.43, 3.74]**	-0.09 [-0.31, 0.12]	**3.77 [3.68, 3.85]**	**5.12 [4.91, 5.33]**
	No Obfuscation	**5.23 [5.08, 5.37]**	-	**3.61 [3.52, 3.68]**	-	-
People in background	Silhouette	-2.11*** [-2.25, -1.96]	1.35*** [1.19, 1.51]	1.80*** [1.61, 1.99]	0.57***[0.48, 0.65]	-0.38*** [-0.61, -0.16]
	3D avatar	-1.01*** [-1.15, -0.86]	0.46*** [0.26, 0.66]	0.97*** [0.78, 1.16]	0.36***[0.26, 0.46]	-1.25*** [-1.51, -0.99]
	Blurring	-1.37*** [-1.52, -1.23]	**4.62 [4.42, 4.82]**	0.72*** [0.52, 0.91]	**4.21 [4.11, 4.31]**	**4.88 [4.64, 5.11]**
	No Obfuscation	**5.96 [5.84, 6.07]**	-	**2.69 [2.52, 2.87]**	-	-

- Mean value index of perceived information sufficiency, perceived privacy, perceived fairness and willingness to share videos on a scale of 1 to 7
- $^{+}$p<.1, * p<.05, ** p<.01, *** p<.001

A.6 Code Systems for Non-expert Evaluation

Table 8: Code systems of qualitative analysis of fairness perception

Categories	Codes	Subcodes	Description of codes	Concerned obfuscated regions
Biases based on test-takers' characteristics	Gender		Participants believed that the gender of the subject could still be inferred from obfuscated regions	Face, body, people in background
		Gender\Body shape	Gender was believed to be inferred from displayed body shape	Body, people in background
		Gender\Attire	Gender was believed to be inferred from displayed attire	Body, people in background
		Gender\Hair	Gender was believed to be inferred from displayed hair length	Face, people in background
	Ethnicity		Participants believed that the ethnicity of the subject could still be inferred from obfuscated regions	Face, body, people in background
		Ethnicity\Skin tone	Ethnicity was believed to be inferred from displayed skin tone	Face, body, people in background
		Ethnicity\Attire	Ethnicity was believed to be inferred from displayed attire	Body, people in background
		Ethnicity\Hair	Ethnicity was believed to be inferred from displayed hair color, type	Body, people in background
	Social status		Participants believed that the social status of the subject could still be inferred from obfuscated regions	Body, background, people in background
		Social status\Background	Social status was believed to be inferred from displayed background	Background, people in background
		Social status\Attire	Social status was believed to be inferred from displayed attire	Body, people in background
Biases based on other aspects	Unprofessionalism	Test-taking environment	Participants mentioned unusual test-taking place and people appearing in the background as unprofessional	Background, people in background
	Distraction		Participants believed unusual test-taking place and presence of others as distracting during review	Background, people in background
		Distraction\Obfuscation design	Attractive design and presence of technical glitch while obfuscation is applied could cause distraction from reviewing	Face, body, background, people in background

A.7 Diagnostic Assessment of Normality and Homoskedasticity Assumptions for LMEMs

Table 9: Shapiro-Wilk Test for Normality and Breusch–Pagan Test for Homoskedasticity of LMEM Residuals

ROI	Dependent variables	Shapiro-Wilk results for normality test (W-statistics, z-score, p-value)	Breusch–Pagan test results for heteroskedasticity
Face	Perceived privacy	W=0.99, z=2.79, p=.002	$\chi^2(1)$=10.75, p<.001
	Perceived info. sufficiency	W=0.95, z=12.32, p<.001	$\chi^2(1)$=13.16, p<.001
	Perceived fairness	W=0.98, z=12.13, p<.001	$\chi^2(1)$=72.85, p<.001
	Composite score	W=0.99, z=2.86, p=.002	$\chi^2(1)$=2.38, p=.12
	Willingness to video sharing	W=0.99, z=0.78, **p=.21**	$\chi^2(1)$=2.36, p=.13
Body	Perceived privacy	W=0.98, z=4.51, p<.001	$\chi^2(1)$=19.11, p<.001
	Perceived info. sufficiency	W=0.96, z=11.99, p<.001	$\chi^2(1)$=8.24, p<.005
	Perceived fairness	W=0.99, z=8.68, p=.01	$\chi^2(1)$=3.73, p=.05
	Composite score	W=0.99, z=1.02, **p=.15**	$\chi^2(1)$=0.59, p=.44
	Willingness to video sharing	W=0.99, z=0.68, **p=.25**	$\chi^2(1)$=0.21, p=.65
Background	Perceived privacy	W=0.96, z=6.37, p<.001	$\chi^2(1)$=0.39, p=.53
	Perceived info. sufficiency	W=0.99, z=7.99, p<.001	$\chi^2(1)$=100.58, p<.001
	Perceived fairness	W=0.97, z=10.04, p<.001	$\chi^2(1)$=29.01 , p<.001
	Composite score	W=0.99, z=0.51, **p=.31**	$\chi^2(1)$=0.06, p=.81
	Willingness to video sharing	W=0.95, z=6.78, p<.001	$\chi^2(1)$=76.32, p<.001
People in background	Perceived privacy	W=0.98, z=5.38, p<.001	$\chi^2(1)$=21.03, p<.001
	Perceived info. sufficiency	W=0.99, z=10.81, p<.001	$\chi^2(1)$=9.75, p<.001
	Perceived fairness	W=0.99, z=8.59, p<.001	$\chi^2(1)$=9.69, p<.001
	Composite score	W=0.99, z=3.75, p<.001	$\chi^2(1)$=30.57, p<.001
	Willingness to video sharing	W=0.99, z=1.36, **p=.09**	$\chi^2(1)$=5.11, p=.02

B Materials used in Expert Evaluation

B.1 Tasks During Experts' Interviews

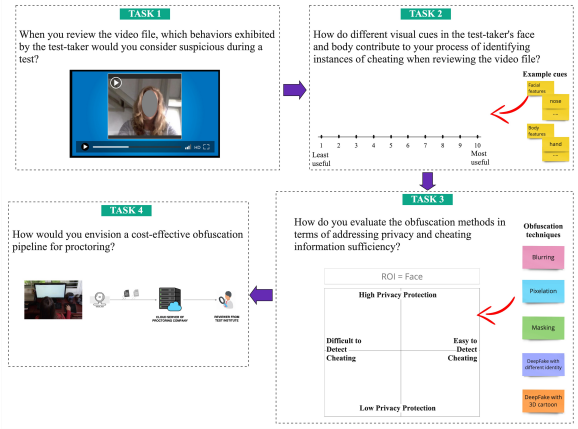

Figure 5: The expert interviews followed a sequence of tasks: *Task 1* involved listing cheating instances for each region in the image shown; *Task 2* focused on identifying and rating (1-10) visual information that could be suppressed through obfuscation without compromising cheating detection. Example only shows cues for face and body. Other regions were also discussed; *Task 3* allowed experts to assess the privacy-cheating trade-off for each method in each region. They were required to drag and drop according to their assessment. *Task 4* involved a discussion on the obfuscation pipeline, guided by the existing flow diagram of remote proctoring

B.2 Recruitment of Experts for Interviews

Table 10: Details of experts being interviewed

Experts	Designation	Expertise
E-Assessment Experts	Programme leader	4+ years in proctoring university exams and privacy related research
	Programme manager	~5 years in proctoring university exams
	Assessment specialist	1+ year in proctoring university exams
Computer Vision Experts	Key researcher	20+ years in CV research, with 10+ years particularly in surveillance domain
	Delivery head	10+ years in the surveillance domain with 5 years in privacy protection research
	Staff research scientist	4 years in digital face manipulation research using neural networks
	Scientific staff	3 years in CV research with biometric data protection
Privacy Experts	Assistant professor	5 years in usable privacy research in remote proctoring
	Associate professor	8 years in HCI research, specializing in usable privacy for ubiquitous systems

B.3 Codebook for Qualitative Analysis of Expert Interviews

(1) **Identify cheating behaviors**: This category has 4 subcategories based on 4 video regions or ROIs.
- 1.1 Cheating behaviors in face
- 1.2 Cheating behaviors in body
- 1.3 Cheating behaviors in background
- 1.4 Cheating behaviors in people in the background

The codes are repetitive for all subcategories, mentioned with example quotes. They are: 1) **dishonest behaviors** *("...it's clear that that this person is talking to someone else")*, 2) **unauthorized behaviors** *("obviously if another person appears on the screen..., I assume the prerequisite for this exam was to be alone in the room")*, 3) **unusual behavior pattern** *("If the shoulder is moving..., the hands are moving, more than needed for typing ends")* and 4) **risk of false flags** *("frozen screen and stuff like that. And so the things can go wrong there...")*

(2) **Useful visual cues for cheating detection**: This category also has 4 subcategories based on 4 video regions or ROIs.
- 1.1 Visual cues in face
- 1.2 Visual cues in body
- 1.3 Visual cues in background
- 1.4 Visual cues in people in the background

Repetitive codes for all subcategories are mentioned with example quotes: 1) **crucial cues to preserve** *("face pose is a logical indicator, or say meaningful to understand where she's looking")*, 2) **traits can be concealed** *("not sure umm...how she looks adds any value to cheating, all it matters to form facial expressions ...that can be helpful to have")*, and 3) **cues as frequent indicators** *("depends on cheating, I think, hmm. most visible part is face ... you can have idea looking at her face if she's cooking something")*

(3) **Evaluation of obfuscation methods**: This category also has 4 subcategories based on 4 video regions or ROIs.
- 1.1 Evaluation with face
- 1.2 Evaluation with body
- 1.3 Evaluation with background

- 1.4 Evaluation with people in the background

Repetitive codes for all subcategories are mentioned with example quotes: 1) **privacy evaluation** *("...blurring is actually low privacy and the actual subject face can be retained with a different filter.. they can see the body and the face, and the colour, probably gender...")*, and 2) **cheating detection evaluation** *("I see you want to change the outfits. Replacement is a good technique in my opinion for privacy, how about they hide cheat sheet in their.. Let's say shirt's pockets. You miss out on important information for your cheating assessment"*)

(4) **Ensure fairness in cheating detection**: This category has 2 subcategories.
   - 1.1 Measures for test-takers: The codes: 1) **gender** *("Doesn't matter if you're woman or a man. Same face for everyone if you are replacing their faces... But Very boring as a reviewer to watch them having the similar face. ok. Not sure about any new bias for it. ...People always try to make an assumption.")*, and 2) **skincolor** *("I would also fix the colour for also to avoid biases when I will put a. Let's say 3D fictional face. I would just keep it like a standard colour umm or something you know so it doesn't give out"*)
   - 1.2 Measures for people in the background: Only one code: **gender** *("...I would prefer virtual avatars. I see. for the full body. It should be same avatar for all ...Keeping gender constant")*

(5) **Practical obfuscation pipeline** This category has two codes: 1) **access to original video** *("The first line of review, the Proctor slash reviewers. They will, they will watch some sort of Anonymized version. Can still detect all the important features. if the secretaries of the examination boards in our situation would watch the normal video without anonymization.")*, and 2) **issues with obfuscated videos** *("I think students would feel much more comfortable. Secretaries can still check if you are wrongly flagged I think. ...huge issues when you rely only on Anonymized versions. Solving privacy issues umm introducing ethical issues even more, some trust issues")*