User Study Protocol for Guardian

Ofek Sela, Noor Atieh, Gali Maman

September 2023

1 Introduction

1.1 background

Guardian is a web application designed to empower users with the ability to immunize their face images against AI manipulations and evaluate the effectiveness of different immunization modes against various manipulation models. Users can upload a face image, choose from multiple immunization modes (no immunization, encoder attack, or diffusion attack), and then test their immunized images against one or two AI manipulation models (Diffusion or StyleGan). The app is asking the user to pick a target image when using diffusion attack, To allow users to customize the defense strategy.

1.2 Objectives

The primary objectives of this user study are as follows:
To evaluate the effectiveness of different immunization modes (encoder attack, diffusion attack) on face images. To assess the impact of AI manipulation models (Diffusion and StyleGan) on immunized face images. To gather user feedback and preferences regarding the app's usability and user experience. We also want to asses the user's sense of security to post face images online after he was introduced to our app.

2 Participant Recruitment

2.1 Participant Criteria

Participants for this study will be recruited based on the following criteria:
Age: 18-40
Technical Expertise: not required

2.2 Recruitment Method

Participants will be recruited through online platforms, social media, and email invitations.

3 Study Design

3.1 Experimental Setup

The user study will be conducted on zoom interviews. Participants can access the Zoom meeting via links that will be sent to them through Email.

3.2 Task Description

Participants will be asked to perform the following tasks using Guardian:
Upload a face image.
Select one or more immunization modes (no immunization, encoder attack, or diffusion attack).
Choose one or two AI manipulation models (Diffusion or StyleGan). If the diffusion attack mode is chosen, customize the defense by selecting a target image.
Click the "RUN" button to initiate image immunization and manipulation. Receive and compare the results of different immunizations and AI models.
Later on, users will be asked a few questions (see Appendix for interview script).

4 Variables and Measures

4.0.1 Dependent Variables

The following dependent variables will be measured:
User satisfaction with the app.
Image quality before and after immunization.
Effectiveness of different immunization modes.
Impact of AI manipulation models on immunized images.

4.1 Independent Variables

The independent variables include:
Immunization modes (no immunization, encoder attack, diffusion attack).
AI manipulation models (Diffusion and StyleGan).
Target images (for diffusion attack)

5 Procedure

5.0.1 Informed Consent

Participants will be required to provide informed consent before participating in the study. They will be informed of their rights and the option to withdraw at any time.

5.0.2 Pre-Study Questionnaire

Participants will complete a pre-study questionnaire, providing demographic information and baseline data.

5.0.3 User Tasks

Participants will follow step-by-step instructions provided by the interviewer to complete the designated tasks.

5.0.4 Data Collection

Data will be collected through user feedback.

6 Data Analysis

6.0.1 Data Preprocessing

Data will be preprocessed, including data cleaning and handling missing values.

6.0.2 Statistical Analysis

Statistical analysis will be performed using scikit-learn and Matplotlib.

7 Ethical Considerations

7.0.1 Privacy and Data Protection

Participant data will be handled with strict privacy measures. Data will be anonymized and stored securely.

7.0.2 Debriefing

Participants will be debriefed about the study's purpose and outcomes upon completion.

8 Reporting and Results

8.0.1 Reporting Format

The results will be reported in a report presentation.

8.0.2 Expected Results

We anticipate that the study will provide insights into the effectiveness of different immunization methods on AI manipulation models, as well as user preferences for customization. Speciffically, we think that the immunizations methods would be less effective on StyleGan.
We expect that letting users to pick their own target images should lead to a more effective immunization.
We assume that awarness to the fact that current immunization modes are not generic, i.e suited for specific AI models, will decrease user's sense of security and his motivation to use Immunization methods.

9 References

1. A. Hertz, R. Mokady, and Y. Nitzan. Protecting against StyleGan-based image manipulation. Technical report, Tel Aviv University, 2022.
2. H. Salman, A. Khaddaj, G. Leclerc, A. Ilyas, and A. Madry. Raising the cost of malicious ai-powered image editing. arXiv preprint, 2023.

10 Appendix - interview script

Thank you for participating in our study! We are going to ask you to rate from 1 to 10 how much the following sentences represent your feelings:

1. I tend to post my photos online.
2. I would feel safe if my images were online.
3. I am aware of the possibility to manipulate my images using AI models.
4. I take actions to protect my images.

Now, we ask you to choose an image in which we can see your face clearly and then, go to our web app (link). On the index page, you need to upload the image from your home computer. After you do, click "Get started". In the next page you are asked to mask your image. please mark the critical areas of your image, i.e the areas which the AI model should not edit (usually the face itself). When you finish, click "Submit" to navigate into the settings page. In this page you are asked to pick one or more immunization modes and one or more AI models. You can see small "?" boxes next to all the fields and options, click on them for Guidance. If you pick multiple options from either immunization modes or AI models, the app will run all the combinations for you to

compare the results. In this part of the interview, we ask you to pick Diffusion model only. For picking "Diffusion attack" immunization, you also need to pick a "custom target image", the immunization method will trick the AI model to manipulate the image into something similar as possible to the target. we will ask you to run the app at least 10 times with "diffusion attack" on at least 5 different images. for each image, we want you to run the app with the grey target, and then run it with other target that you choose. It's recommended to choose target with colours that are contrasted to your original face image. Note that it's also possible for you to upload target image from your own computer. Now I'll give you 20 minutes to play with the app and try the different features. When you finish, please rank the following sentences:

5. I wouldn't mind using a protected image instead of the original one.
6. I can't notice the difference between original photo and immunized one when using encoder attack.
7. The difference between original photo and immunized one when using encoder attack is a fair price for my security.
8. I can't note the difference between original photo and immunized one when using diffusion attack.
9. The difference between original photo and immunized one when using diffusion attack with grey target is a fair price for my security.
10. The difference between original photo and immunized one when using diffusion attack with custom target is a fair price for my security.
11. I feel safe posting the original photo.
12. I feel safe posting the protected photo which was immunized by encoder attack
13. I feel safe posting the protected photo which was immunized by diffusion attack with grey target.
14. I feel safe posting the protected photo which was immunized by diffusion attack with custom target.
15. I would use the app again to immunize my images.

Great, now we get to the final part of the interview, where you need to do the same as before, but now please use StyleGan model as well. When you finish, please rank the following sentences:

16. I wouldn't mind using a protected image instead of the original one.
17. The difference between original photo and immunized one when using encoder attack is a fair price for my security.
18. The difference between original photo and immunized one when using diffusion attack with grey target is a fair price for my security.
19. The difference between original photo and immunized one when using diffusion attack with custom target is a fair price for my security.
20. I feel safe posting the protected photo which was immunized by encoder attack.
21. I feel safe posting the protected photo which was immunized by diffusion

attack with grey target.
22. I feel safe posting the protected photo which was immunized by diffusion attack with custom target.
23. I would use the app again to immunize my images.