The Subversive AI Acceptance Scale (SAIA-8): A Scale to Measure User Acceptance of AI-Generated, Privacy-Enhancing Image Modifications

JACOB LOGAS, Georgia Institute of Technology, USA
POOJITA GARG, Georgia Institute of Technology, USA
ROSA I. ARRIAGA, Georgia Institute of Technology, USA
SAUVIK DAS*, Carnegie Mellon University, USA

To resist government and corporate use of facial recognition to surveil users through their personal images, researchers have created privacy-enhancing image filters that use adversarial machine learning. These "subversive AI" (SAI) image filters aim to defend users from facial recognition by distorting personal images in ways that are barely noticeable to humans but confusing to computer vision algorithms. SAI filters are limited, however, by the lack of rigorous user evaluation that assess their acceptability. We addressed this limitation by creating and validating a scale to measure user acceptance — the SAIA-8. In a three-step process, we apply a mixed-methods approach that closely adhered to best practices for scale creation and validation in measurement theory. Initially, to understand the factors that influence user acceptance of SAI filter outputs, we interviewed 15 participants. Interviewees disliked extant SAI filter outputs because of a perceived lack of usefulness and conflicts with their desired self-presentation. Using insights and statements from the interviews, we generated 106 potential items for the scale. Employing an iterative refinement and validation process with 245 participants from Prolific, we arrived at the SAIA-8 scale: a set of eight items that capture user acceptability of privacy-enhancing perturbations to personal images, and that can aid in benchmarking and prioritizing user acceptability when developing and evaluating new SAI filters.

CCS Concepts: • **Security and privacy → Social aspects of security and privacy**; *Privacy protections*; • **General and reference → Measurement**; • **Human-centered computing → Heuristic evaluations**; User studies.

Additional Key Words and Phrases: scale development, subversive artificial intelligence, human factors, acceptability, mixed methods, measurement theory, adversarial machine learning

ACM Reference Format:
Jacob Logas, Poojita Garg, Rosa I. Arriaga, and Sauvik Das. 2024. The Subversive AI Acceptance Scale (SAIA-8): A Scale to Measure User Acceptance of AI-Generated, Privacy-Enhancing Image Modifications. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 185 (April 2024), 43 pages. https://doi.org/10.1145/3641024

*Also with Georgia Institute of Technology, School of Interactive Computing.

Authors' addresses: Jacob Logas, logasja@gatech.edu, Georgia Institute of Technology, School of Interactive Computing, Atlanta, Georgia, USA; Poojita Garg, poojita.garg@gatech.edu, Georgia Institute of Technology, Atlanta, Georgia, USA; Rosa I. Arriaga, arriaga@cc.gatech.edu, Georgia Institute of Technology, School of Interactive Computing, Atlanta, Georgia, USA; Sauvik Das, sauvik@cmu.edu, Carnegie Mellon University, Human-Computer Interaction Institute, Pittsburgh, Pennsylvania, USA.

2573-0142/2024/4-ART185
https://doi.org/10.1145/3641024

1 INTRODUCTION

Advances in computer vision (CV) have enabled algorithmic surveillance at an unprecedented scale, both by governments and corporations [39, 86, 140]. In response, researchers have explored using these same advances to subvert computer-vision based surveillance by creating privacy-enhancing image filters through the use of adversarial machine learning (AML) (e.g., [24, 25, 115]). AML encompasses a suite of techniques to create "adversarial examples" of inputs into machine learning models that aim to look nearly identical to people but cause the model to mis-classify the input (evasion attacks) or learn false associations during training (poisoning attacks) [71, 83, 107]. These properties present an opportunity for improving image privacy in the face of CV-based surveillance. Today, an emerging class of privacy-enhancing image filters use AML to perturb images in a manner that minimally alters the image while aiming to thwart automated detection and recognition (e.g., [24, 25, 115]). Prior work coins the phrase "Subversive AI" (SAI) to describe this use of AML to disrupt surveillance of end-users by institutional actors [31].

Fawkes, one of the earliest SAI projects, was met with much public interest; it was featured in the New York Times and has been downloaded nearly a million times [61]. The broad appeal of Fawkes poses a unique opportunity to increase user adoption of privacy-enhancing technologies. Indeed, several more SAI filters have been developed and/or published to build on this success (e.g., [24, 24, 25, 82, 105, 114, 133, 141]). Yet, to date, these SAI filters have primarily been evaluated on technical attack efficacy (e.g., how effectively they cause commercial computer-vision systems to misclassify adversarial inputs), and have not been systematically evaluated with the human users they claim to protect.

Decades of prior work in usable privacy and security suggest that privacy and security technologies designed without centering end-user awareness, motivation, and ability are unlikely to see sustained use or widespread adoption [15, 32, 33, 77, 103, 109]. Moreover, prior art in the broader CSCW, computer-mediated communication (CMC), and HCI literature is clear that self-presentation is a primary concern for people when sharing personal information online [35, 108, 119]. Thus, one hypothesis is that without consideration of user preferences in developing SAI filters, user acceptance of the privacy-enhancing perturbations introduced by these filters will be low. Indeed, SAI filters approximate user acceptance by optimizing for metrics that are easy to measure without user input, like pixel distance — the smaller the change made, the better for user acceptance. Prior work suggests, however, that user acceptance of novel technologies is multi-faceted[32, 36, 122] and, thus, unlikely to be captured solely by these simple metrics [112, 113]. Absent a standard measure for user acceptance of privacy-enhancing filter perturbations, SAI researchers face a large cost to answer seemingly simple questions like: *"Do people find the perturbations introduced by Fawkes acceptable for images they want to share online?"* To address this challenge, we introduce a validated psychometric scale to measure user acceptance of privacy-enhancing image perturbations: the subversive AI assessment scale, or SAIA-8.

Psychometric scales "measure [behavioral] phenomena that cannot be observed directly." [37] Indeed, the HCI literature contains many measures that aid in bridging the gap between quantitatively focused research and human factors [65, 69, 95, 125, 134]. Following best practices from measurement theory literature[7, 62], we employed an inductive approach to scale development that spans three steps:*construct exploration*, *item generation*, and *item refinement*.

In the first step, *construct exploration*, we qualitatively investigated factors that contribute to privacy-enhancing image filter acceptance. We interviewed $N = 15$ participants to gather candid reactions to SAI filters and more naive privacy enhancing approaches (e.g., blurring faces). Overall, participants reacted negatively to the SAI perturbations and shared skepticism toward the claimed privacy benefits. Our exploration and analysis of SAI filter reactions uncovered four factors that

contributed to interviewee acceptance of SAI filters — aesthetics, identity modification, shareability, and skepticism of protection. We also found that the goal of "minimal pertubration" is counter-productive for end-user acceptance. When perturbations made by SAI filters were *not* clearly visible, users were not convinced of the effectiveness of the filter at protecting their privacy. When the perturbations *were visible*, users often disliked the changes made to their images, describing the outputs as having an "uncanny valley" effect.

In the second step, *item generation*, we generated $Q = 106$ potential items from a combination of qualitative analysis insights and paraphrased statements from participants, distributed across the four factors from our qualitative analysis. The potential items were presented to $N = 47$ participants with survey experience for feedback on *content validity*, allowing us to reduce down to a set of $Q = 53$ high-quality items.

In the third step, *item refinement*, we arrived at the final eight items through iterative application of measurement theory principles [7]. We performed four sequential iterations of refinement wherein we solicited responses from, in total, $N = 215$ participants on the Prolific study participation platform[1]. Each iteration was analyzed on the basis of item heuristics: difficulty, discrimination, reliability, correlation, and validity [1]. The final SAIA-8 scale (see Table 2) achieved a Cronbach $\alpha = 0.87$.

To validate that our final scale measured acceptance, we conducted a *convergent validity* analysis with $N = 30$ participants [22]. This analysis investigated the degree to which the scale adhered to proxies of acceptance we identified from the interviews. We found that the SAIA-8 strongly correlates with the proxy for *identity modification*, and weakly correlates with the proxies for *aesthetics* and *shareability*, but not with the proxy for *skepticism of protection* — an expected result, since we removed items related to general privacy concerns in the final scale. Accordingly, the SAIA-8 can be viewed as a measurement of user acceptability of privacy-enhancing image perturbations based on three factors: aesthetics, identity modification, and shareability. Moreover, it can be non-redundantly paired with a scale to measure general privacy concerns and attitudes (e.g., the IUIPC [88]).

In summary, our work makes the following contributions:

- We introduce, to our knowledge, the first validated psychometric scale measuring user acceptability of privacy-enhancing image perturbations: the SAIA-8.
- We build and extend prior literature analyzing how people feel about different privacy-enhancing image filters, particularly emphasizing newer techniques powered by adversarial machine learning.
- Based on our interview data and responses to the SAIA-8, we distill key design insights and implications for creating more user acceptable privacy-enhancing image filters.

2 RELATED WORKS

Image privacy has long been an area of interest in social computing. Subversive AI is the latest in a long tradition of prior work seeking to help users obscure, encrypt, or otherwise protect sensitive information in personal images (image privacy-enhancing technologies, or image PETs) [57, 84, 85]. To date, SAI technologies have primarily been developed optimizing for technical attack efficacy and human imperceptibility [31, 116]. This lack of focus on user-needs, in turn, likely negatively affects widespread user adoption [32]. Part of the challenge is that there remains no standard measurement for human-acceptability that is applicable to privacy perturbation tools, making it difficult and expensive to evaluate SAI filters from a human-centered perspective. We review the

[1]https://prolific.co

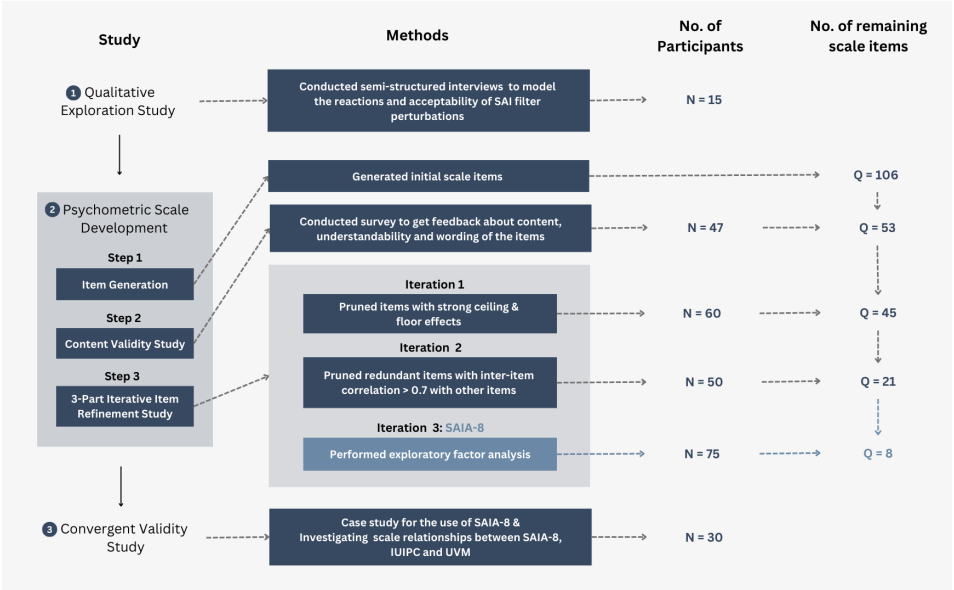


Fig. 1. Schematic diagram for the research flow

prior literature on image privacy-enhancing technologies and the role of psychometric scales in human-centered design, and discuss how our work builds on and extends these threads.

2.1 Image Privacy-Enhancing Technologies

The rise of ubiquitous surveillance and facial recognition models motivated the research and development of image PETs [131]. These tools aim to give users greater autonomy over the distribution and use of their images. In the past, this goal was achieved through use of well-known and understood processes – e.g., steganography, encryption, obstruction, nudges [3, 20, 30, 49, 56, 57, 85, 100, 102, 135]. However, few image PETs gained widespread adoption outside of privacy-focused users thanks in part to their usability and perceived usefulness [29, 46, 73, 109].

In the last few years, researchers have approached the problem of institutional privacy surveillance by appropriating AML techniques [24, 25, 82, 105, 114, 115, 133, 141]. These techniques claim to provide protection from algorithmic observation by perturbing pixels in an image to disrupt model inference [107, 123]. Aside from the primary goal of disrupting algorithmic surveillance, these models are fine-tuned to favor "imperceptibility" as AML practitioners have done in the past. "Imperceptibility" is often quantified through Pixel p-Norm, Earth Mover's Distance, Structural Similarity Index, or Deep Net Embedding [112]. Sen et al. challenged the appropriateness of using these heuristics to capture human perception, finding them to be insufficient approximations of human perception [113]. While work continues to develop a metric that better approximates human perception[50], the literature fails to challenge this primary assumption. These measures, while not explicitly stated as such, serve as a hypothesized proxy for human acceptability — i.e., perturbation perception is assumed to negatively correlate with human acceptance. We challenge this hypothesis, confirming through our investigation that acceptability for SAI filters is a multi-faceted construct that goes beyond the visibility of perturbations [36, 122]. Moreover, we design and validate the first scale, to our knowledge, that aims to measure the human acceptability of image perturbations made to improve privacy.

2.2 Human-Centered Evaluation

The HCI community has long held a role in computer science research to examine and explain technological advancements through a sociotechnical perspective [33, 89, 106, 139]. Indeed, ML and security advancements have struggled with acceptance due in part to a lack of consideration for human factors [11, 34, 91]. The usable security and privacy community has a running character – "Johnny" – for their evaluations of security advancements [15, 48, 60, 96, 109, 132]. Technologies like encryption or cryptographic signatures could serve to greatly improve user security and privacy, but such technologies are often considered a hindrance by users [10, 47, 98, 130]. In recent years, the HCI community took a greater interest and role in the design and evaluation of human-centered privacy tools [27, 41, 80, 85]. Relatedly, ML applications are often viewed as "black boxes" with little to no intuitive description associated with the decisions they make [23, 43, 81, 138]. Thus, the adoption of ML within image PETs creates a unique intersection of acceptability challenges from both disciplines, making existent acceptance measures insufficient. Our work explores this intersection of acceptability, accounting for factors in acceptance which illustrate a need for human-centered design in the development of privacy-enhancing image perturbations.

However, integration of human factors into SAI filter research would require extensive experimentation. Fortunately, psychometric scales can capture difficult-to-observe behavior in a way that can easily integrate into development pipelines [7, 16, 51, 53, 62, 64]. Indeed, the HCI literature contains many behavioral measures that aid in bridging the gap between quantitatively-focused research and human factors [65, 69, 95, 125, 134]. In the realm of privacy and security, psychometric scales can determine privacy concerns[17, 88], security intentions[42], level of value for others' privacy[55], and security attitudes[44]. Efforts to quantify behavior in psychometric scales serve to ease the integration of human factors into highly technical systems. The SAIA-8 scale continues this tradition by giving researchers interested in privacy-enhancing image perturbations a low-cost method of integrating a human perspective into their development pipeline.

2.3 Factors in Image Sharing

There are a variety of complex factors that contribute to a persons' motivation to share personal information online [9, 94, 127]. Prior literature has investigated how one's preferred presentation of self influences sharing on social platforms [35, 45, 66, 101, 111]. Additionally, work has explored how a person's ideas about privacy can influence sharing behaviors [4, 139]. Also, social pressures have been found to contribute to sharing amongst several groups [13, 28, 52, 54, 68, 124]. These behaviors are further complicated by the introduction and use of computational photography and aesthetic image filters found on contemporary social media sites [12, 70, 72, 99, 129]. Here, we contribute to the literature by exploring how factors in image sharing could be altered through the use of an image privacy enhancing technology.

3 STUDY 1: QUALITATIVE EXPLORATION

The first step in psychometric scale development is the creation of questionnaire items either through deductive or inductive means [62]. The deductive approach involves generating scale items based on prior work which provides a strong theoretical foundation for the construct of interest — in our case, human acceptability of anti-surveillance perturbations made to images. The inductive approach involves a ground-up approach to item generation based on inferences made from analyzing qualitative data sourced from, e.g., interviews or prior literature. We employed the inductive approach because, to the best of our knowledge, there is no widely accepted theory for what drives human acceptability of privacy-enhancing image modifications. Without this theory, there would be no strong basis to deductively generate candidate items for inclusion in the scale.

We conducted semi-structured interviews with $N = 15$ participants to model their reactions to and acceptability of privacy-enhancing perturbations. In our analysis, we identified four themes which we reference for item generation: *Aesthetics*, *Identity Modification*, *Shareability*, and *Skepticism of Protection*. *Aesthetics* captures participants' reactions to the visual composition of the image — e.g., color, clarity, noise. *Identity Modification* captures participants' feelings on how the filter altered the subjects' identity presentation — e.g., health, age, gender. *Shareability* captures participants' preferences towards sharing the filtered image with an intended audience. Finally, *Skepticism of Protection* captures participants' perceptions of how the filtered output addresses their privacy concerns and whether they believe they provide privacy protection. Based on these four themes and the interview transcripts, we generated $Q = 106$ candidate scale items formatted as statements with a 5-point Likert agreement response. [21, 92]

3.1 Data Collection

We collected candid reactions to a set of existing anti-surveillance image filters with hour-long semi-structured interviews. The goal of these interviews was to probe participants' reactions to the perturbations created by the filters. We asked participants to apply four different anti-surveillance filters to their own personal images: two popular SAIs filters (Fawkes [115] and LowKey[25]) and two non-SAI (face pixelator, emoji stickers). While the filters we chose varied in approach and threat model, the overarching goal was the same — to improve users' online privacy against algorithmic surveillance by perturbing pixels on and around a subject's face. We transcribed and analyzed interview transcripts using open coding.[110] We grouped the codes resulting from this analysis into themes for item generation.

3.2 Recruitment

We employed multiple recruitment strategies to find a diverse and representative sample of the general population for this study: social media posts, flyers posted locally, the Prolific participation platform[2], and snowball sampling. Each advertisement included a simple description of the study — e.g., "Give your opinion on privacy enhancing technologies!" — and listed the eligibility requirements for the study (e.g., over 18, resides in U.S.) as well as the offered compensation ($20). This recruitment campaign resulted in $N = 15$ interview participants — summarized in Table 1 — who met eligibility requirements and fully completed the interview study.

3.3 Interview Protocol

We approach the participant interviews with the following research questions:

- What threats do users wish to protect themselves against when using privacy-enhancing image filters? Why?
- How do users feel about the changes made to their personal images?
- What are the factors that influence the *shareability* of the filtered image?
- What are the factors that influence users' *preference* for one filter over another?

We interviewed participants individually, guiding them through a webapp — developed by the research team — where they could apply each SAI filter to a personal photo. Interviews were performed and recorded with the Zoom conferencing software and transcribed using the Otter.ai[3] automated transcription service with the participants' informed consent. We assume participants would have a more visceral and grounded reaction to their own images being modified due to prior work in self-presentation. [35, 108, 119] Therefore, we asked participants to bring five personal

[2]https://prolific.co
[3]https://otter.ai

Participant	Ethnicity	Hispanic	Age Range	Gender	Education
1	Black	No	18-24	Transgender	Bachelor's degree
2	Black	No	25-34	Man	Bachelor's degree
3	White	No	25-34	Transgender	Bachelor's degree
4	Asian	No	18-24	Man	Bachelor's degree
5	Asian	No	35-44	Woman	Graduate or professional degree
6	Black	No	18-24	Man	High school diploma or GED
7	White	No	35-44	Man	Graduate or professional degree
8	Black	No	25-34	Man	Associates or technical degree
9	Black	No	25-34	Woman	Bachelor's degree
10	Black	No	25-34	Man	Some college
11	Black	No	25-34	Woman	Bachelor's degree
12	Black	No	25-34	Woman	Bachelor's degree
13	Asian	No	18-24	Man	Bachelor's degree
14	Asian	No	25-34	Man	Bachelor's degree
15	White	No	25-34	Prefer not to say	Graduate or professional degree

Table 1. Demographic information for the 15 participants in Study 1.

images to the interview that they "considered sharing but did not because of privacy concerns." As we were requesting potentially sensitive content that participants were uncomfortable publicly sharing, **we did not collect or view the images and assured participants that their images would not be collected.** Our interview protocol followed three sequential steps: *orienting*, *perturbation reaction*, and *preference*.

In the first part, *orienting*, we asked participants questions about their current privacy behaviors — e.g., "Do you use privacy protecting tools such as incognito browsing, a virtual private network, or locked social media accounts?" We also asked participants to critically reflect on what incited the adoption of these behaviors and how often they engaged in them. We wanted to actively prime participants to think about their current privacy behaviors because, in pilot studies, we found that participants often underestimated their adherence to privacy practices. These pilot participants often assumed a nonchalant attitude toward privacy which on further investigation oftentimes did not reflect their behavior. As our goal was to understand factors that drove the acceptability of image perturbations that were introduced specifically for privacy purposes, we wanted participants to be primed to think about their personal privacy concerns. Once introduced, participants more readily considered the benefits the filter could bring.

We next asked participants about the five personal photos they were asked to bring to the interview. Participants described their hesitation with sharing the selected photos publicly. We further prompted them to describe tools or behaviors they may use to protect their privacy if they were to share the photo with friends or family. These questions were used to give the interviewer insight into the participant's privacy attitudes toward personal images, helping to contextualize the reactions they had to the filter perturbations.

In the next phase, *perturbation reaction*, we had participants apply the SAI and non-SAI filters to the personal photos they had brought to the interview, investigating their candid reactions to the changes made. Perturbations were applied through a web application developed by the research team and hosted on Google Cloud. We developed a web application for each of the image filters we tested using the Gradio library for Python. [2] The host was equipped with an NVIDIA T4 GPU to accelerate filters that supported such optimizations. Participants applied each of the filters to

one of their personal photos — the one most aligning to institutional privacy concerns.[4] Once the filter was applied, we asked participants about their immediate reaction to the perturbed image including the differences they noticed and their affective response to them. Then, we asked how those changes made them feel about sharing the photo on social media and whether the promise of improved privacy impacted their willingness to share the image. If the filter had multiple levels of protection (e.g., Fawkes has a low, medium, and high mode), we asked which one the participant would be more comfortable sharing publicly and why. Finally, we asked how participants would feel using the filter on the last image of themselves that they shared on social media. We repeated this process for each of the four filters in turn.

After the participants experienced all four filters, we asked them to comparatively evaluate them. We requested participants to rank the filters they saw and give their reason for the rankings, asking for elaboration on how contextual factors — e.g., where the photo would be posted, what the photo contains — influenced their reasoning. Finally, the researcher called back to the information divulged in the orientation to determine if other privacy concerns might affect the ranking.

3.4 Qualitative Coding

The inductive approach to questionnaire development involves generating candidate items based on emergent themes from qualitative data [62]. We employed open coding to analyze the interview transcripts, resulting in a corpus of unordered codes – the code-book.[110] Open coding involves identifying snippets of the conversation that relate to the research questions and classifying them into one or more codes.[110] This is an iterative process that often requires many passes over the data and is typically performed in parallel by independent researchers who compare and discuss their individual code-books. Here, two researchers performed three passes over the interview transcripts; each time refining the code-book. Our first pass had two researchers independently develop a code-book by going over the entire corpus of transcripts. These independently developed code-books were then compared and disagreements over the classifications were resolved through discussion and analysis of the transcript snippets relating to the code or codes. We then took the combined code-book back to a subset of the data — those transcripts that the researchers identified as having rich data. This pass confirmed that existing codes were appropriate and had researchers record any emergent codes. The second pass was similarly scrutinized by both researchers, discussing codes between each other and refining the code-book by pruning, merging, or renaming codes. The final pass over the data confirmed the appropriateness of the codes and resulted in no new codes for discussion; however, after this pass the researchers still discussed the current code-book and further refined the code-book by combining codes we agreed to be similar to each other.

3.5 Findings

Our analysis of the interviews surfaced four contributing factors for the participants' reactions to and acceptability of the SAI filters we presented: **Aesthetics**, **Identity Modification**, **Shareability**, and **Skepticism of Protection**. These themes provided us with an empirical grounding for generating candidate items for the SAIA scale. Here, we will speak of each overall theme as it relates to the acceptability of the privacy filters.

3.5.1 Aesthetics: "I just don't like what it did to my face". Several interview participants found the perturbations made by the filters to negatively impact the **aesthetic** of the image. Filter alterations could at times generate discordant artifacts or alter color composition.

[4]In pilot studies the filters would occasionally not work with a participant's photo, hence we requested users to bring five photos to avoid this problem. However, none of the participant photos in the main study encountered this issue.

Filters also had a tendency to place artifacts on the subject's face in aesthetically unpleasing ways. Participant 4 expressed the filter "doesn't really blend well with the rest of the image" due to the filter applying a "mask" to the face. Elaborating that "the skin shade on the rest of the body is very different" and the "skin texture is very different." Participant 4's concern was with how the SAI filter inconsistently applied perturbations across the image. Thus, aesthetic harmony between the parts of the image that are and are not perturbed appears to be an important consideration for user acceptance of these filters.

Color changes could affect participants' skin tone — "My face is beet red for some reason" (P15) — or alter the perceived lighting in the photograph — "Makes me look like a dude at a rave" (P18). Most participants noted how LowKey modified the color of the image. Participant 2 felt the filter made it seem like "something has been painted on the face." In these cases, and others, the color changes were viewed as a clear negative. As P18 stated: "I prefer the low [filter mode] solely because i don't like the coloration patterns that are used overall." However, we also found evidence that color changes can be acceptable or preferred — "I like medium better than low, almost like a cartoon" (P3) or "If the coloration pattern was different, like purple and green" (P18). Thus, some users appeared to be open to color changes if they were more in-line with their aesthetic preferences or if they felt more in control.

From these findings, we brainstormed $Q = 25$ candidate items for inclusion in the SAIA scale, including:

- "I feel the filtered image is less clear."
- "I don't feel comfortable with the changes made to the photograph."
- "The filter worsens the color in the photo."

3.5.2 Identity Modification: "They distort the face just enough that it makes you uncomfortable". Participants also noted that the perturbations made by SAI filters affected a subjects' expressed **identity** presentation along four dimensions: *age*, *health*, *gender*, *humanity*.

In several cases, participants mentioned that the Fawkes filter altered the subjects' perceptible *age*. Participants described the LowKey filter as applying changes that made the subject appear "more mature,"(P2) with one remarking "it looks like it aged me a bit"(P3). Participant 5, for example, noted the artifacts introduced by LowKey made them look "kind of weird and like, wrinkled."

The filters also gave the appearance of poor *health*. Participant 15 remarked that LowKey "makes me look diseased for some reason." Other participants described images filtered by Fawkes as "like I got sick or something"(P3), "imposing a skin disease on somebody"(P7), and "like I've been in the sun too long"(P7). Several participants went beyond describing the changes as simply imposing an illness, saying, for example, that Fawkes "just makes me look like a zombie"(P18).

Participants — especially those who identified as queer — also found that the perturbations could affect the subjects' gender expression. Participant 18 was not pleased with how the LowKey filter modified their gender presentation: "There are some people who would probably love you know, like obfuscation that also includes a free mustache. I am not one of them. In fact, the photo has me and like, probably 40 minutes of makeup, specifically to avoid that possibility." In contrast, Participant 3 — a transgender woman who was actively transitioning — found the changes made by both Fawkes and LowKey as affirming to her identity: "because I haven't had facial feminization yet, this one's extremely validating because my brow ridge is gone, my cheeks are where they're supposed to be, and my jawline actually has the right contours. I'm pretty pleased." Of the 15 participants, P3 was the only one to have a positive reaction to their filtered image and expressed a willingness — if not a desire — to use the filters in the future.

Finally, some participants felt the filters dehumanized or objectified the subject. This reaction mainly arose from the naive approach which obscures the face for privacy enhancement – see

Figure 2. This perception of "dehumanizing" the subject was surprisingly inconsistent among the participants with some thinking the high and/or emoji setting was dehumanizing while others felt the 'medium' setting was more dehumanizing. For example, Participant 2 found the high setting to be "too much, the face is blurred too much." This participant preferred the low pixelation because "at least [with low] I can see the face a bit, that its a human." On the other hand, Participant 7 preferred the maximum pixelation and/or emoji option because "it's like you put a piece of paper over his face, [it] seems less dehumanizing [because] I can imagine removing the sticker over his face." Participant 18 had a preference for "medium [pixelation] because it doesn't obscure what the thing is to the human eye." They chose medium over low due to their idea that the low could be more easily reversed by an advanced algorithm.

Through these insights, we generated $Q = 30$ candidate items to capture respondent attitude toward how filter perturbations change the photo subjects' identity presentation:

- "The filter is affirming to my gender identity."
- "I look older after the filter was applied."
- "The filter makes me look less human."

3.5.3 Shareability: "Probably hiding my face is a better way". Participants found that the perturbations introduced affected the **shareability** of the images because the perturbations compromised the *purpose of sharing* the image or influenced the *expected reaction of the viewer*.

Prior work has shown that there are several reasons for a person to share images online[94]: e.g., seeking and showcasing experiences, social connection with close relationships, and reaching out to a wider audience. Participants stated that the changes introduced by the filters could at times defeat the purpose of sharing the image altogether. For example, P4 discussed a photo illustrating an accomplishment that they would potentially want to share with a community with common interests, but felt that full face obfuscation would make sharing the photo significantly less appealing: "It's an achievement to move or get healthy, I'd probably prefer to have my face with that." P7 saw sharing photos as a means of building community; using filters made them feel "like my life was a little less sincere and authentic" because it "manipulates my actual appearance too much." P18 felt the filters detract from their sharing of photos of themselves because "the photos [of myself] exist to be identifiable... I am not a raccoon" in reference to the emoji filter. Moreover, when discussing applying the filter to a photo P3 had already shared, they mentioned "that picture was sentimental with a lot of emotion behind it" and they felt the filter would detract from the sentimentality.

There were also situations where participants found that the filters *increased* the shareability of images. P18, who mentioned they would not want to use filters that fully obfuscated their own face, said: "I can see situations where I'd apply [a filter] to pictures of other people in frame" and "a friend always publishes photos of their kids with emojis which I think is adorable." Similarly, P15 felt as though the emoji filter could be used in a way that appropriately retains the purpose of sharing the image while providing improved privacy protections: "Part of the reason why people have their face in images is to show their personality. So blocking your face out with an emoji, I think still accomplished 80% of what you'd want a portrait for in a casual setting."

Another factor influencing the shareability of the filtered image was how interviewees expected others in their social circle to react to the image. Participant 15 mentioned that "even if a ML model would have difficulty recognizing it, I would feel uncomfortable with people I know seeing it." Similarly, Participant 4 had a preference for the naive approaches as "hiding my face a better way, [Fawkes and LowKey] look like I'm giving a false representation of myself." The subtle but noticeable perturbations made by the AML-powered filters could also conflict with self-presentation goals. As P18 jested,"I could just be a horrifically zombie-like creature and chosen a really bad

headshot." Finally, P3 was concerned that the perturbations introduced would draw unwanted attention and expressed apprehension about having "to explain to friends and family why I'm putting up this [filtered] person" and that "it'd be risky to open up about why it is important to me to be private as a trans person."

Based on these findings, we developed $Q = 19$ candidate items for inclusion in the SAIA scale, including:

- "The changes made by the filter defeat the purpose of sharing the image."
- "I wouldn't share the image publicly after the filter was applied."
- "My family or friends would ask about the filtered photo if I posted it on social media."

3.5.4 Skepticism of Protection:"If a friend knew me, they would still be able to point that out". Many participants also exhibited skepticism about the level of protection provided by the filters. This **skepticism of protection** arose from two main factors: *recognizability* and *visible markers of protection.*

The first factor contributing to skepticism with the efficacy of the filters was that the AML-powered filters maintained the recognizability of the subject. Because the protections provided were somewhat abstract and non-intuitive, participants had a hard time believing that they were effective. One participant indicated that "if I wanted to be private, I would probably [use a naive approach] because you can't make out any facial features or anything" (P5). Generally, participants felt as though the filters did not do enough to mask the subjects' identity as "someone who knows me would be able to tell that it's me" (P15) or because the subjects' "body is still showing" (P2) revealing tattoos, birthmarks, or other identifying features. Participant 10 succinctly described why he doesn't believe the protection "because I can still see who the person is."

Others were willing to accept that subtle perturbations could provide privacy, however they generally felt more comfortable with obvious obfuscation. "Without that additional information for certainty, I have to say I like the naive protection application simply because it's the only one where I know what it's doing" (P18). There was a further "worry that visual cues alone might give people a misleading impression of the actual degree of obfuscation provided, because it could be that the [Fawkes model on] low is doing something even though I can't see a difference" (P18). In cases where privacy was paramount such as in "some super secret chat" a participant felt most comfortable "just using the one where my face is really pixelated" (P5).

While the purpose of the AML-powered filters is to maintain human-legibility of subjects while protecting against algorithmic classification, these findings echo the need for what Do et al. call *perceptible assurance* in providing privacy protections: i.e., the need for people to intuitively understand and perceive, for themselves, how a privacy-enhancing technology can practically protect one's privacy. [41]

From these findings, we generated $Q = 32$ candidate items to capture people's belief in the effectiveness of the privacy protections promised, including:

- "I am confident that applying this filter would protect my privacy."
- "The filter addresses my privacy concerns for the photo."
- "I could do better to increase my privacy than the filter."

Our process of semi-structured interviews and qualitative analysis resulted in the four themes described: **aesthetics**, **identity modification**, **shareability**, and **skepticism of protection**. These themes in turn informed the development of the SAIA-8 scale.

4 STUDY 2: PSYCHOMETRIC SCALE DEVELOPMENT

Building off insights from our interview study, we next followed a multi-step process to develop a scale that captures human acceptability of privacy-enhancing perturbations made to personal

images, spanning *item generation*, *content validity*, *iterative item refinement*, and *convergent validity*. Through this process, we refined our set of candidate items to the final $Q = 8$ items.

4.1 Item Generation

We used the findings from the qualitative exploration study to generate $Q = 106$ potential items for the scale. Generated items were statements to which respondents could provide a level of agreement along a 5-point Likert scale ranging from Strongly Agree to Strongly Disagree [21, 92] Two researchers worked independently to brainstorm statements and paraphrase quotations by interview participants related to the emergent themes described above. The resulting items were at times highly similar with one another, but we elected to keep similar items for the initial refinement iterations if we agreed the wording was sufficiently different — e.g., "My face is the same as it is in the original image." and "The filtered image looks like me." We grouped generated items into the four high-level themes that were used to generate them: aesthetics, identity modification, shareability, and skepticism of protection. The full initial set of items generated can be found in Appendix C.

4.2 Content Validity Study

Content Validity refers to the extent to which the candidate items appropriately represent the psychometric construct we would like to measure: i.e., human acceptability of privacy-enhancing perturbations made to personal images. [8, 118] Proceeding with the inductive method of psychometric scale development[62], we took steps to refine the $Q = 106$ statements down to a small set of internally consistent items.

We performed an initial item reduction and refinement based on *content validity* using feedback we elicited from participants who were experienced in questionnaire development and participation. The goal of this analysis is to ensure that items in a scale are relevant to the construct of interest, capture all relevant content, and are high-quality. Both expert evaluation and pilot testing are traditionally used to aid in improving content validity. [8, 118] In this study, we used a mixed-methods approach to evaluate the initial set of $Q = 106$ questionnaire items, pruning and refining them and thus, resulting in an overall set of $Q = 53$ for further refinement. This study involved $N = 47$ participants in total.

4.2.1 Survey Design. The survey was implemented as a web form hosted on Qualtrics (Appendix D).

Participants were presented with a pre-filtered version of one of 6 unique profile pictures sourced from the Pexels stock photography site.[5] Each of the 6 profile photos were preprocessed with the pixelation mask, Fawkes[115], and LowKey[25] filters (see Figure 2). We elected to use preprocessed images because we wanted to do the initial filtering with the broadest possible set of users. Requiring participants to upload and filter their own photos during this phase of the process would have potentially biased our sample towards those who both had a high-quality face picture readily available, were willing to upload that photo, and were willing to wait up to 15 minutes for the filter to be applied.

In the survey instructions, we directed participants to imagine themselves as the photo subject and to provide any feedback on the items they encounter including feedback about the content, understandability, or wording of the items. Subsequent pages of the survey presented a single candidate item with both an original image and a preprocessed output displayed. The interface allowed images to be clicked, opening a modal with the selected image scaled to 75% of the window width. Each participant was asked to view and respond to 10 candidate items, giving free-response

[5] https://www.pexels.com/license/

(a) $Pixelate_L$ (b) $Pixelate_M$ (c) $Pixelate_H$

(d) $Emoji$ (e) $Fawkes$ (f) $Lowkey$

Fig. 2. Example of filter outputs.

feedback in a text box. Participants were given no more than 5 questions per page so the photos and survey items were visible at the same time on the majority of devices.

4.2.2 Recruitment. We performed a content validity analysis with experts and participants experienced in responding to surveys. The experts we recruited were 17 of our academic colleagues and non-academic acquaintances who had survey development expertise and no prior knowledge of the project. Our sample of 30 experienced survey respondents were recruited from Prolific, with an approval rate of 75% and at least completed 100 studies. Participants were required to be above the age of 18 and reside in the United States. Participation took on average 10 minutes for which participants were compensated $5.

Participants were majority women at 65% and 29% men with the remaining 6% identifying as non-binary or third-gender. Most respondents in this iteration identified as Caucasian (83%) with the rest of the participants identifying as either Asian (3%), Black (3%) or mixed race (9%). Most of the participants had received higher education, including associate (6%), some college (26%), bachelor's degree (29%), or graduate (19%), while 16% had completed high school and 3% had not.

4.2.3 Analysis and Refinement. Our analysis followed a feedback-first approach: i.e., we focused on identifying feedback that referred to the format, understandability, relevance, or described a strong affective response to the statement and not the filtered image. Through this approach, we identified items that might be of low *content validity*. Items we identified were further examined and discussed amongst two researchers, either leading to refinement of the language or pruning.

On occasion, we also added new items to the set of candidates based on participant feedback. In general, the items we added were to provide an inverse to existing potential items — e.g., "I am

happy with the filtered output" had its inverse "I am unhappy with the filtered output" appended. We found 16 items to lack appropriate inverse statements in the set of potential items, thus $Q = 16$ new items were added.

The feedback we collected also directed us to rephrase several items. We decided to rephrase the item if participants expressed confusion over what was being asked, like with Q46:

> **Q46**: The changes made by this filter are jarring.
> "What's jarring?"
> "Unclear on the definition of jarring in this context."

When generating items, we at times made use of colloquialisms or obscure vocabulary that confused participants. Here, "jarring" was an unfamiliar term and thus we rephrased it to "The changes made by this filter are surprising."

Through this process, we filtered our candidate set down to $Q = 53$ items that had high content validity.

4.3 3-Part Iterative Item Refinement Study

We refined our scale over three more iterations of data collection and analysis. Following best practices outlined and followed in prior literature[1, 42, 62], we collected responses to the remaining candidate items after each iteration. We pruned and refined candidate items along the following five dimensions of interest: **item difficulty**, **item discrimination**, **inter-item correlation**, **internal consistency**, and **factor analysis**. We describe each, in turn, below.

Item Difficulty Item difficulty refers to the level of difficulty or ease with which respondents answer an item [1]. Traditionally, it is used as a measure for "difficulty" of an item where there is an objectively correct answer – i.e., academic tests. As our items are Likert format and have no objectively correct response, item difficulty is analyzed by computing the percentage of respondents per possible response. These distributions of responses along the scale of Strongly Agree to Strongly Disagree are analyzed for strong ceiling or floor effects. Low variability in these items make it difficult to distinguish between individuals or groups who have different levels of ability or traits.

Item Discrimination While item difficulty tells us the average level of performance on items, item discrimination evaluates how well items discriminate between respondents with high and low levels of the construct measured [1]. Item discrimination is measured as the point-biserial correlation coefficient which measures the relationship between item score and the total test score for individuals who selected the response versus those who did not.

Items with a high point-biserial correlation coefficient are generally candidates for retention because they can reliably differentiate between individuals who score high and low on the overall construct. Additionally, high discrimination items tend to have less measurement error which can contribute to a higher reliability for the scale as a whole.

Inter-Item Correlation Inter-Item Correlation measures the homogeneity between items in a scale — i.e., the correlation strength between one item and other items in the scale [1]. In general, we want moderate levels of inter-item correlation: too low, and the items may not measure the same construct at all; too high, and the items may be redundant. Measurement theory literature suggests item correlation range between $\rho = 0.3$ and $\rho = 0.7$ [1].

Internal Consistency Internal consistency refers to the extent to which the items in a scale measure the same construct [1]. Cronbach's alpha is one commonly used measure of internal consistency in psychometric scale development. It measures how well the items are related to each other and is calculated as the average of all possible split-half correlations.

A high Cronbach's alpha indicates the items in a scale likely measure the same construct. Measurement theory literature suggests psychometric scales to have internal consistency over $\alpha = 0.7$ [1].

Factor Analysis Factor analysis is a statistical method that aims to identify a small number of latent variables among a set of correlated, observable variables — e.g., participant responses to a psychometric scale. In scale development, the goal of factor analysis is to identify item groups that represent the same underlying latent variables. With knowledge of these item groups, we can further filter down candidate items by retaining only a parsimonious subset of items from each group.

There are two approaches to factor analysis: confirmatory (CFA) and exploratory (EFA). CFA is a deductive approach in which the analyst expresses a hypothesis of the number of latent variables in a dataset, and which items correspond to those latent variables, and statistically tests how well the observed data conform to that hypothesis [6]. EFA is an inductive approach in which the analyst starts with no pre-conceived hypothesis and aims to identify emergent latent variables in the observed dataset [128].

The first and second iteration focused on filtering and refining based on *Item Difficulty*, *Item Discrimination*, and *Inter-Item Correlation*. The third iteration used the previous three methods along with *Internal Consistency* and *Exploratory Factor Analysis* to arrive at the final scale. This process gave us the SAIA-8 scale consisting of 8 final items, with an *Internal Consistency* of $\alpha = 0.87$ which contains latent factors associated to *aesthetics*, *identity modification*, and *shareability*.

4.3.1 Recruitment. We recruited 60, 50, and 75 participants respectively for each iteration on the Prolific platform. Participants were required to be 18 years or older and reside in the United States. We selected the participants for each iteration so each candidate item was viewed and responded to at least 15 times for each filter we tested.

Participants over all three iterations were majority men with 55% with women (41%) making up most of the remainder. We had 3% non-binary/third-gender respondents and 1% who preferred to not disclose their gender identity. Like our content validity recruitment, most participants were college educated: associates (10%), some college (20%), bachelor's (44%), graduate (12%). Several participants, however, completed high-school (13%), did not complete high-school (1%), and others preferring not to disclose (1%).

4.3.2 Iteration 1 & 2. In iteration 1, we refined the set of candidate items from $Q = 53$ to $Q = 45$ items. For this iteration, our analysis largely focused on dropping items with high item difficulty — i.e., we removed items if responses to those items showed a high ceiling or floor effect. Q14, for example, was removed: "I would like to use this filter even if it didn't protect my privacy."; among the 60 participants in this iteration, 30 answered Strongly Disagree with the remainder either choosing Disagree or Neither Agree nor Disagree. Figure 3 illustrates items removed for their item difficulty.

In the second iteration, we refined the set of candidate items from $Q = 45$ to $Q = 21$ items. For this iteration, we identified and removed redundant items with an inter-item correlation above $\rho = 0.7$. As an example, Q95: "I am happy with the filtered image" was strongly correlated with Q54: "I feel comfortable with publicly posting the filtered photo online, if it improves my privacy" and Q102: "I don't feel comfortable with the changes made to the photograph." Initially, Q54 was dropped due to its 0.7 correlation with Q95, low discriminatory power, and overloading with both feelings about aesthetics and skepticism of protection. We then decided to drop Q95 in favor of Q102 due to Q95 having lower discriminatory power and its wording being non-specific.

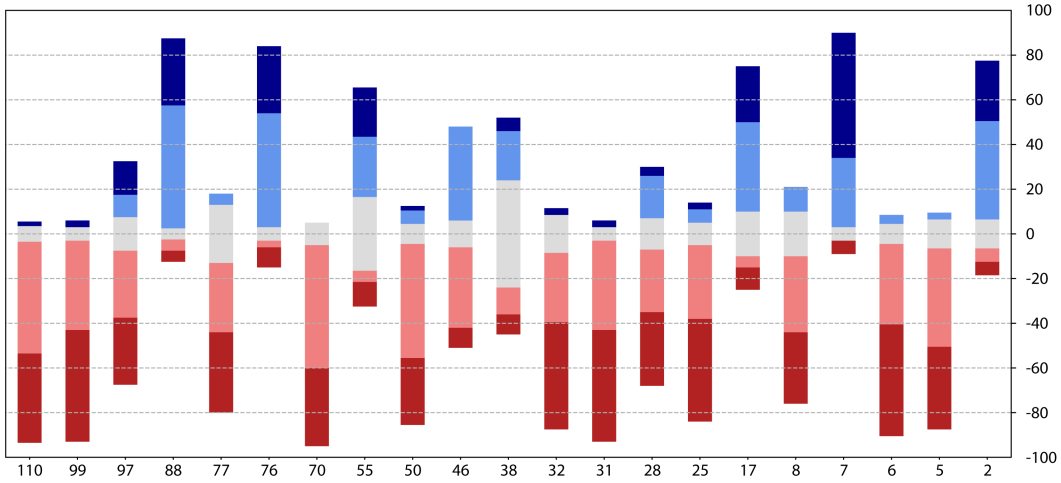


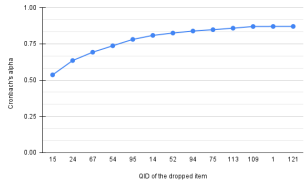
Fig. 3. Candidate items pruned after the first iteration and the normalized distribution of responses.

4.3.3 Iteration 3: SAIA-8. In the third iteration, we reduced the remaining $Q = 21$ items to the final set of $Q = 8$ items that make up the SAIA scale. Beyond the pruning criteria for iterations 2 and 3, we performed EFA on the items. Candidate items that loaded weakly onto the emergent latent variables in the scale were further analyzed for pruning, taking into account the criteria used in the prior iterations. Finally, we considered how the *Internal Consistency* was affected by selective removal of remaining items.

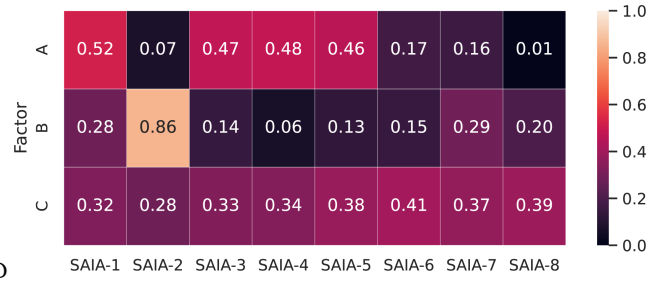
We performed Principal Component Analysis (PCA) with 2, 3, and 4 factors with an eye toward the categorization we previously associated with each candidate item: aesthetics, identity modification, shareability, and skepticism of protection. If an item was weakly correlated to the latent variable to which it was loaded, we reasoned why this may be and dropped the item if we agreed the item heuristics proved it to be weaker than the other items to which it was similar. We pruned five items this way, leaving 16 candidate items. However, the overall *Internal Consistency* was still unacceptably low — $\alpha = 0.44$.

To improve internal consistency, we iteratively experimented with pruning items to assess their impact on *Internal Consistency*. Through this process, we pruned 8 more items to reach our final set of $Q = 8$, increasing α from 0.44 to 0.87. It is important to note that *Internal Consistency* is just one heuristic for reliability and as such simply pruning based only on the α score would be irresponsible. Rather, when we identified an item that a had a strong negative impact on consistency, the research team discussed the item as a candidate for pruning, incorporating other measures of item performance as necessary.

Through this process we pruned all of the items relating to participants' skepticism of protection. Doing so greatly increased α, but the decision to drop these items as a group was also driven by agreement that beliefs about the efficacy of these filters — while important — was exogenous to the specific perturbations made and thus orthogonal to the other candidate items in the scale. The other items — which measured aesthetic preference and willingness to share the perturbed image — were more specific to perturbations made by the filters. Moreover, in our interview study, we found that that participants' perceptions with respect to the effectiveness of the promised privacy protections were a less a significant factor for acceptability than aesthetic and sharing preference.



(a) Cronbach's alpha versus QID of the dropped item



(b) 3 factor PCA loading for the final 8 items.

In following the inductive approach to psychometric metric development[62], we arrived at a final scale with 8 items and an internal reliability of $\alpha = 0.87$. The final set of 8 items are listed in Table 2.

Items 1, 3, 4, and 5 – factor 'A' in Fig 4b – relate to the aesthetic of the filtered image, capturing the users' reaction to how it *looks*. Computational photography literature suggests there are several factors that exist within the umbrella of image aesthetics [76, 87]. However, with the SAIA-8, we were not concerned with measuring each aesthetic factor individually, as other measures exist for this purpose [126]. Instead, our goal was to capture the user's subjective feeling of employing the filter on their images.

Item 2 – factor 'B' in Fig 4b – captures the user's reaction to the *identity modification* that occurs when applying a filter. Recall that a user's preferred presentation is vital to what they decide to share online [35, 108, 119]. Our goal was to capture the user's feeling that the filter has contributed to or deviates from their preferred presentation.

The last three items, factor 'C' in Fig 4b, relate to users' *willingness to share* these filtered images. Once again, there are several factors that are attributed to why a person may or may not want to share an image, as well as attempts to develop scales to measure these factors [38, 52, 67, 119]. Our goal was to capture a user's willingness or likelihood to share the filtered image with others.

ID	Item	Factor	μ	σ
1	I don't feel comfortable with the changes made to the photograph.	A	3.53	1.17
2	The filter makes me look less human.	B	3.83	1.23
3	I feel concerned with how the filter has affected my looks.	A	3.47	1.22
4	I feel the filtered photo's changes are immediately noticeable.	A	4.13	1.11
5	My family or friends would ask about the filtered photo if I posted it on social media.	A	4.33	1.21
6	The changes made by the filter defeat the purpose of sharing the image.	C	3.87	1.25
7	I wouldn't share the image publicly after the filter was applied.	C	3.93	1.34
8	I would rarely use this filter for photos shared to social media.	C	4.17	1.21

Table 2. The SAIA-8 questionnaire items with the primary factors identified by PCA along with the mean (μ) and standard deviation (σ) of responses in the convergent validity study. Participants answer each item on a 5-point Likert scale ranging from Strongly Disagree (1) to Strongly Agree (5).

5 STUDY 3: CONVERGENT VALIDITY STUDY

In our final study, we performed a **convergent validity** analysis with $N = 30$ participants to compare the SAIA-8 measure with alternative measures of the construct or principles of the construct—referred to as *proxies* [22]. Convergent validity provides evidence to judge the construct

validity of a measure in measurement theory [93]. Properly performing convergent validity requires choosing proxies that align with properties of the construct under measurement – i.e., acceptability. Testing the convergent validity of a new scale against appropriate proxies is an important step in assessing if the construct that the scale is measuring abides by theory-driven expectations. As Carlson et al. state, "weak correspondence provides less certainty that data actually reflect the properties of the intended constructs"[22]. However, perfectly convergent proxies are rare. Measurement theory literature suggests convergence between $\rho = 0.50$ and $\rho = 0.70$, with a preference for convergence above that range and avoidance for convergence below it [7]. However, because the SAIA-8 scale sits at the nexus of many theoretically orthogonal constructs (image aesthetics, self-presentation / identity, shareability, and privacy concern), we expected values at the low-end of that range.

5.1 Approach

We asked participants to apply either the LowKey or Fawkes — on the "High" setting — filters to personal photographs and respond to the SAIA-8 scale along with four proxy scales. We selected a proxy scale for concepts we found affected the acceptability of privacy-enhancing perturbations: *skepticism of privacy*, *identity modification*, *aesthetics*, and *shareability*. Participation time was estimated at 15 minutes and each participant was compensated $3.50 upon completion of the survey. As we asked participants to apply these filters to their own photos, **we assured participants that images would not be retained or used for any other reason**. The experiment description in Prolific and on the first page of the survey expressed "We can neither see the photo you upload nor will we be storing it anywhere." The study protocol as described was approved by the IRB.

For the aesthetic proxy, we chose a measure of image aesthetics described by Keelan[76] with response options set by Siahaan[117]. For the identity modification proxy, we chose a measure of the Uncanny Valley Effect (UVE) as described by Ho [63]. For the shareability proxy, we chose a scale capturing motivations of photo sharing [94]. Finally, for the skepticism of privacy proxy, we included the Risk Beliefs, Collection, and Unauthorized Secondary Use portions of the Internet Users' Information Privacy Concern (IUIPC) scale [88]. The proxy surveys are provided in Appendix F.

The survey was partially randomized with SAIA-8 always being presented first and the proxies presented in a random order afterward. Each set of items corresponding to SAIA-8 or the proxies were presented in a uniformly random order. Deviating from the previous survey design, the aesthetic[76, 117] and shareability[94] proxies were measured for both the original image and the filtered image independently. We reasoned that the image content — which we did not control — could account for a low aesthetics or shareability score independent of perturbations made by the filter. Accordingly, our analysis considered the delta between the aesthetics and shareability items for the original and filtered images so that we could isolate the impacts made by the filters.

We performed correlative analysis on the survey responses by $N = 30$ participants. Some proxies called for certain items to be reverse-coded for analysis [63, 88]. We used the *Spearman* ρ for correlation measurement between the scales due to its appropriateness for Likert scale responses [90]. Our analysis had two stages: *inter-scale correlation* and *inter-item correlation*.

In our *inter-scale correlation* analysis, we sought to calculate how the scales compared against each other *as a whole*. The scales were condensed to a single numerical value — a score — for each participant response: specifically, we summed the individual items of the SAIA-8 and each of the proxies. We then calculated the *Spearman* ρ between each scale sum. We found ρ values within or above the acceptable range between the SAIA-8 scale and each of the proxies. However, condensing the SAIA-8 and its proxies to a single score may have obscured some co-factors within the proxies.

As measures oftentimes contain co-factors, we also performed an *inter-item correlation* analysis. This analysis provided insight into the how each of the SAIA-8 items aligned with individual items in the proxy scales. We found several items from the SAIA-8 scale that captured constructs from other scales individually that the *inter-scale* analysis did not capture. Overall, we found the SAIA-8 scale adequately correlated — $\rho >= 0.50$ — with proxy scales, suggesting it has strong convergent validity with the construct of acceptance.

5.2 Recruitment

The $N = 30$ participants we recruited from Prolific had to be 18 years of age or older and to reside in the United States. Our participants mostly identified as white (90%) with one mixed race (3%) and two self-describing as "Other" (7%). The gender identity distribution was more diverse, with 48% identifying as men, 45% as women, and 7% as non-binary / third gender. Most of our participants made less than $50,000 per year (44%) with 30% making between $50,000 and $100,000. The remainder made over $100,000 (23%) or preferred not to answer (3%). Finally, all of our participants had a high-school education at minimum with a majority either in college (37%) or finishing with a degree (33%). A few completed a graduate degree 7% or an associates degree 7%.

5.3 Results

The results from our *inter-scale* and *inter-item* analyses illustrate SAIA-8 has acceptable correlation with these varying validated proxies. Our proxy for identity modification was the most highly correlated in both analyses with $\rho = 0.61$ inter-scale correlation with SAIA-8.

For the inter-scale analysis, we calculated the *Spearman* correlation between all the scales (see Figure 5). The SAIA-8 scale was well correlated with the identity modification proxy ($\rho = 0.61$) while the aesthetics and shareability proxies were weakly correlated — $\rho = 0.47$ and $\rho = 0.43$ respectively. The proxy we chose for skepticism of privacy, on the other hand, had little correlation with $\rho = 0.12$. This result was expected considering our prior decision to remove items that captured concerns related to the privacy effectiveness of the filters. Although the aesthetics proxy is below 0.5, it was more highly correlated with SAIA-8 than any of the other three other scales. The shareability to use proxy was similarly correlated to skepticism of protection ($\rho = 0.45$) and SAIA-8 ($\rho = 0.43$). This result may indicate that the shareability proxy is capturing distinct, orthogonal constructs to the SAIA-8 and external factor.

Fig. 5. Spearman ρ of the selected measures' total score—Internet Privacy Attitudes (IPA), Uncanny Valley Effect (UVE), Aesthetics (AES), Shareability (SH)

The *inter-item* analysis revealed greater insights into the relationships between these scales (see Figure 6). Overall, the SAIA-8 scale saw its greatest correlation with the items in the UVE scale. The second statement in particular "The filter makes me look less human" was correlated with 11 of the proxy items, especially the semantic scale between "Synthetic" and "Real." Also of note is the correlation between the third SAIA item and "Crude:Stylish" on the UVE measure at $\rho = 0.73$. The last UVE correlation of interest is the "Plain:Weird" factor which had a $\rho >= 0.7$ for statement 4, 5, 6, 7, and 8 of SAIA-8.

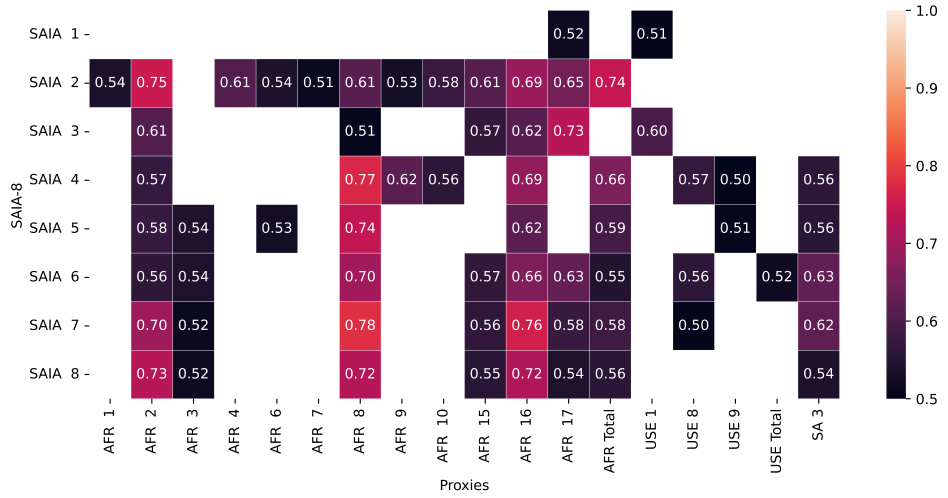


Fig. 6. Items across the proxies with correlation greater than $\rho = 0.5$

The aesthetic scale (AES) also saw stronger correlations with specific items on the SAIA-8 scale. The first factor, "Sharpness" was somewhat correlated with the first ($\rho = 0.51$) and third ($\rho = 0.6$) SAIA-8 items. The other correlative factors reflect more about the changes to the subject with AES 8 and 9. AES 8 correlated with items 4 ($\rho = 0.57$), 6 ($\rho = 0.56$), and 7 ($\rho = 0.50$). AES 9 was also correlated with item 4 ($\rho = 0.50$) but additionally correlated with item 5 ($\rho = 0.51$).

One statement from the sharing scale (SH) was correlated with the SAIA-8. This item — "To cheer myself up" — correlated across items 4 ($\rho = 0.56$), 5 ($\rho = 0.56$), 6 ($\rho = 0.63$), 7 ($\rho = 0.62$), and 8 ($\rho = 0.54$). Finally, none of the Internet Privacy Attitudes items we tested had a *Spearman* correlation over $\rho = 0.50$.

In summary, our convergent validity analysis showed high convergence with the identity modification proxy, weak convergence with the aesthetics and shareability proxies, and low convergence with the skepticism of privacy proxy. We discuss the implications of these results more in the Discussion (Section 6).

6 DISCUSSION

In this paper, our goal was to arrive at a brief psychometric scale for evaluating privacy-enhancing perturbations — focusing on SAI filters — from a human perspective. Our qualitative exploration with 15 participants found the SAI filters we experimented with — Fawkes and LowKey — to be generally unacceptable and attributes four factors that contribute to this unacceptability. The factors identified were used to generate a large corpus — $Q = 106$ — of potential scale items which we refined over three iterations with $N = 232$ unique participants in total. The final SAIA-8 scale achieved an internal consistency of $\alpha = 0.87$ — well above the suggested $\alpha = 0.70$ from measurement theory. [7] Lastly, we validated the SAIA-8 scale through an investigation of its convergence with proxies that measure factors of acceptance we previously identified — i.e., *content validity*. The content validity experiment found the SAIA-8 scale to appropriately align with the factors of acceptability we identified with one notable exception, skepticism of protection, due to relevant statements being dropped during refinement. In the complete process of developing the SAIA scale, we gained insights for the design of future privacy enhancing technologies.

6.1 Benchmarking human acceptability of SAI filters using the SAIA-8

Benchmarks are important in machine learning research, and can help accelerate the pace of progress as researchers and practitioners compete. To date, however, there have been few benchmarks to capture human acceptability, preferences, and attitudes in the development of AML-powered privacy-enhancing image filters — so perhaps it is unsurprising that the best tools, to date, have not produced outputs that appear to be acceptable to human users. We envision the SAIA-8 scale as a means to benchmark human acceptability of adversarial image filters such that future innovations strive to beat prior approaches not just in technical attack efficacy, but also in human acceptability.

To date, SAI filters are researched, trained, and evaluated by researchers primarily trained in machine learning, privacy, and security. These researchers are often not cross-trained as user experience researchers, making user assessment more challenging and time consuming. To that end, the SAIA-8 provides a standard measurement that we hope will be easy to use in lieu of a more careful study design. Deploying machine learning model outputs to distributed crowdworkers for assessment and annotation is already a familiar process; thus, we expect that it should be relatively simple for those who develop SAI filters to use the SAIA-8 in their development workflows.

In sum, the SAIA-8 provides an important new metric for evaluating and iteratively improving SAI filters in a manner that centers human needs.

6.2 Perceptible assurance and identity affirmation drive acceptability of privacy-enhancing image perturbations

Participants in our interview study generally preferred the naive approach to privacy protection — i.e., pixelation or emoji. The perturbations made by these obfuscating approaches, while more obvious, afforded an intuitive understanding of how their privacy might be protected (i.e., by making the subject's face unrecognizable). In contrast, participants found it harder to understand how and believe that the SAI filters, which have more subtle changes, provided privacy protections. For example, when filters like Fawkes were used on a lower-strength setting where the perturbations were less obvious, participants expressed greater reluctance to use those filters: indeed, they didn't seem to change anything about the image at all, and so the privacy concerns remained. Recent work in tangible privacy affords one explanation as to why: users exhibit greater trust in privacy controls that provide *perceptible assurance* over those that must be blindly trusted [5, 40, 41]. Accordingly, we dropped questions directly related to participants' trust in the effectiveness of privacy protections from the SAIA-8 scale. Our goal in developing the SAIA-8 scale was to capture the acceptability of perturbations made to images for anti-surveillance purposes; while trust is important in driving the acceptance and adoption of any technology, trust can be influenced by any number of exogenous variables beyond the image perturbations introduced by the filters (e.g., media consumption, peer influence [32]).

Importantly, however, our findings do *not* suggest that SAI filters can never be human acceptable: rather, we found that the "imperceptibility" goal is misaligned with acceptance. Approaches like Fawkes and LowKey aim to keep the perturbed image as close to the original as possible, but prior work has shown that this approach is detrimental for privacy goals [104, 112, 113] and our work has shown that this approach is detrimental for user acceptability. But, our work also shows that there are situations where anti-surveillance perturbations can be user acceptable, if not preferable. For example, one of our participants, a transgender woman, enthusiastically accepted the Fawkes and LowKey filters *because* of the perturbations. She found the changes were affirming to her preferred self-presentation. The privacy protections were important but secondary [85]; rather, the filters addressed a more primary concern when sharing content on the social internet: self-presentation. [35, 108]

Relatedly, our convergent validity results suggest that the SAIA-8 scale is most correlated with our proxy scale for identity: i.e., that the factors that drive the acceptability of anti-surveillance image perturbations correlate most strongly with how users responded with how participants viewed the perturbations affected their identity presentation.

To summarize, in our development of the SAIA-8 scale, we identified ways it might be possible to improve user acceptability of SAI filters. First, our data suggest that SAI filters should not be designed with "imperceptible" perturbations as a goal: imperceptibility affords little perceptible assurance and lowers user trust in the effectiveness of the filter. Second, by relaxing the constraint of needing perturbations to be "imperceptible", SAI filters can instead be designed to create perceptible changes that are aesthetically pleasing and identity affirming. Indeed, users of many popular social media applications — e.g., TikTok and Snapchat — enjoy applying image filters that apply stylized effects to their images that they find pleasing, humorous, or enjoyable [12]. Might it be possible to create such filters in a manner that also affords improved privacy protections?

6.3 Evaluating Non-SAI filters using the SAIA-8

While our focus for SAIA-8 was to measure acceptability of SAI filters, there may be benefit to using it to measure user acceptability for other image modification tools as well. Other approaches to image privacy work to obfuscate the users' identity by applying perturbations to the image without using AML — e.g., "cartooning"[56], "stickering"[136]. The factors we identified and used to develop the SAIA-8 scale may well apply to these non-SAI approaches for improved acceptance. It may be that a non-SAI approach can attain greater acceptance while avoiding potential pitfalls of SAI. [104, 105] As an example, two of our interview participants were more accepting of the emoji option over Fawkes or LowKey as it contributed to the subject's character.

6.4 Aligning SAIA-8 with the Technology Acceptance Model (TAM)

The TAM as described by Davis [36] has long been used to understand how and why users accept new technologies. Our inductive approach to the SAIA-8 did not directly make use of the TAM to guide our investigation; however, the factors for acceptance we identified align well with the *perceived usefulness*, *attitude toward using*, and *behavioral intention to use* determinants. Perceived usefulness is described as "the prospective user's subjective probability that using a specific application system will enhance his or her job or life performance." [121] We align this concept of usefulness with our identity modification factor — P3's enthusiastic reaction to the gender affirming changes and others' negative reaction to application of "skin conditions". Attitude toward use is "concerned with the user's evaluation of the desirability of employing an application." [121] We align this with the aesthetic factor — participants' feeling of the changes being immediately noticeable and perhaps raising questions from family or friends. Lastly, the behavioral intention to use factor is "the measure of the likelihood of a person to employ the application." [121] We associate this factor with the shareability factor in the SAIA-8 — participants' feelings like they would use the application to share potentially sensitive images.

7 LIMITATIONS

Our work has limitations that should be considered when interpreting the results. First, our sample was limited to the U.S. specific and not census representative: participants in our item refinement studies, for example, primarily identified as white. However, we did take care to gather a diverse perspective in our studies: specifically, we recruited a diverse group of interviewees across race, education, and gender expression. We did so in large part because prior work suggests that minority populations can sometimes have greater need for effective privacy-enhancing technologies because they disproportionately bear the harms of institutional surveillance. [19, 26, 120] However

as Henrich et al.[58, 59] argues, behavioral studies have a tendency to center on western, educated, industrialized, rich, and democratic populations to make generalizations about worldwide populations. Indeed, investigation finds that privacy attitudes and behaviors vary across cultures. [14, 74, 75, 78, 79, 137, 137] Accordingly, due to the sampling bias in our studies, we may have missed factors that affect acceptability of privacy-enhancing perturbations for populations outside of the United States. In future work, it would be pertinent to see how different populations and cultures vary in responses to the SAIA-8 to identify if there is a need for more population and culture-specific versions of the scale. Nevertheless, given that no such scale existed prior to this work, the SAIA-8 should provide a strong foundation for further hypothesis testing to that end. Another limitation is that the factors we integrate into SAIA-8 do not exhaustively align with all determinants for acceptance expressed in prior literature – i.e., TAM[36, 121, 122]. Namely, we do not capture *external variables* – i.e., sociotechnical factors that influence technology acceptance [32, 36, 97]. Nor do we capture *perceived ease of use* as our protocol controlled for user interface of the filters. However, this measure can be paired with existing measures of *external variables* – e.g., privacy or security attitudes [44, 88] – and *ease of use* – e.g., the SUS [18]. Thus, SAIA-8 can be viewed as a tool for capturing the three determinants of acceptance discussed before – *usefulness*, *attitude toward using*, and *behavioral intention to use* – which previously had no standard scale for privacy perturbations. Given the increasing interest in privacy-enhancing image filters, the SAIA-8 should be helpful for researchers and practitioners working on building and evaluating such tools.

8 CONCLUSION

In this work, we developed and validated a scale to measure user acceptability for privacy-enhancing image filters, in particular those powered by adversarial machine learning. Following the inductive approach to scale development described by Hinkin, [62], we first performed a qualitative investigation of user reactions to existing privacy-enhancing image filters with $N = 15$ interviews wherein participants applied the filters to their own images. Our qualitative analysis identified four factors for filter acceptance *Aesthetics*, *Identity Modification*, *Shareability*, and *Skepticism of Protection*. Using statements by our participants and these factors as guidance, we next generated $Q = 106$ possible items for the SAIA scale. Over three successive iterations evaluated by measurement theory heuristics [7] we refined the survey to $Q = 8$ with a Cronbach alpha of $\alpha = 0.87$. We validated the resulting SAIA-8 measure by comparing it with four *proxies* for factors related to the construct of SAI acceptability. Our findings suggest that users find the perturbations introduced by existing subversive AI image filters to be unappealing and unacceptable, despite their general interest in using such tools for anti-surveillance purposes (as demonstrated by the popularity of tools like Fawkes [61]). The SAI filters often introduced artifacts that many people found aesthetically unpleasing and misaligned with their preferred identity presentation: e.g., participants said the filters made them look sick or less human. Moreover, by aiming to reduce pixel distance between the original and perturbed images, many participants had trouble believing in the effectiveness of the filters at protecting their privacy. However, we also found evidence to suggest that it is possible to design SAI filters in a manner that users may find acceptable by allowing for modifications of the image that are more aesthetically pleasing and identity affirming. In summary, given popular interest in the image filters that subvert algorithmic surveillance, the SAIA-8 provides a way for researchers and practitioners to capture previously difficult-to-measure determinants of privacy-enhancing image filter acceptance. Our hope is that this standard measurement, in turn, will help drive innovation towards more human acceptable solutions.

ACKNOWLEDGMENTS

This work was generously funded, in part, by NSF Grant #2316287. The authors thank Dr. Cori Faklaris for her advice on statistical analysis and evaluation of psychometric measures. The authors also thank the GT SPUD and GT Ubicomp Health and Wellness lab for feedback which helped improve our study design and prototype. Lastly, we would like to thank the reviewers for their valuable reviews.

REFERENCES

[1] 2001. Principles of Test Construction. In *Introduction to measurement theory*. Waveland Press, 118–147.

[2] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ML models in the wild. https://doi.org/10.48550/arXiv.1906.02569

[3] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. 2017. Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online. *ACM Computing Surveys (CSUR)* 50, 3 (Aug. 2017), 44:1–44:41. https://doi.org/10.1145/3054926

[4] Shane Ahern, Dean Eckles, Nathaniel S. Good, Simon King, Mor Naaman, and Rahul Nair. 2007. Over-exposed? privacy patterns and considerations in online and mobile photo sharing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. Association for Computing Machinery, New York, NY, USA, 357–366. https://doi.org/10.1145/1240624.1240683

[5] Imtiaz Ahmad, Taslima Akter, Zachary Buher, Rosta Farzan, Apu Kapadia, and Adam J Lee. 2022. Tangible privacy for smart voice assistants: Bystanders' perceptions of physical device controls. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–31. Publisher: ACM New York, NY, USA.

[6] Sabri Ahmad, Nazleen Nur Ain Zulkurnain, and Fatin Izzati Khairushalimi. 2016. Assessing the Fitness of a Measurement Model Using Confirmatory Factor Analysis (CFA). *International Journal of Innovation and Applied Studies* 17, 1 (July 2016), 159–168. https://www.proquest.com/docview/1807743937/abstract/81957862A4FE4C64PQ/1 Num Pages: 10 Place: Rabat, Morocco Publisher: International Journal of Innovation and Applied Studies.

[7] Mary J Allen and Wendy M Yen. 2001. *Introduction to measurement theory*. Waveland Press.

[8] Enas Almanasreh, Rebekah Moles, and Timothy F. Chen. 2019. Evaluation of methods used for estimating content validity. *Research in Social and Administrative Pharmacy* 15, 2 (Feb. 2019), 214–221. https://doi.org/10.1016/j.sapharm.2018.03.066

[9] Mary Jean Amon, Rakibul Hasan, Kurt Hugenberg, Bennett Bertenthal, and Apu Kapadia. 2019. Influencing Photo Sharing Decisions on Social Media: A Case of Paradoxical Findings. (Sept. 2019).

[10] Reza Anaraky, Bart Knijnenburg, and Marten Risius. 2020. Exacerbating Mindless Compliance: The Danger of Justifications during Privacy Decision Making in the Context of Facebook Applications. *AIS Transactions on Human-Computer Interaction* 12, 2 (June 2020), 70–95. https://doi.org/10.17705/1thci.00129

[11] Hala Assal, Stephanie Hurtado, Ahsan Imran, and Sonia Chiasson. 2015. What's the deal with privacy apps? a comprehensive exploration of user perception and usability. In *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia (MUM '15)*. Association for Computing Machinery, New York, NY, USA, 25–36. https://doi.org/10.1145/2836041.2836044

[12] Saeideh Bakhshi, David Shamma, Lyndon Kennedy, and Eric Gilbert. 2015. Why we filter our photos and how it impacts engagement. In *Proceedings of the International AAAI Conference on Web and social media*, Vol. 9. 12–21. Issue: 1.

[13] Beth T. Bell. 2019. "You take fifty photos, delete forty nine and use one": A qualitative study of adolescent image-sharing practices on social media. *International Journal of Child-Computer Interaction* 20 (June 2019), 64–71. https://doi.org/10.1016/j.ijcci.2019.03.002

[14] Steven Bellman, Eric J. Johnson, Stephen J. Kobrin, and Gerald L. Lohse. 2004. International Differences in Information Privacy Concerns: A Global Survey of Consumers. *The Information Society* 20, 5 (Nov. 2004), 313–324. https://doi.org/10.1080/01972240490507956 Publisher: Routledge _eprint: https://doi.org/10.1080/01972240490507956.

[15] Zinaida Benenson, Gabriele Lenzini, Daniela Oliveira, Simon Parkin, and Sven Uebelacker. 2015. Maybe Poor Johnny Really Cannot Encrypt: The Case for a Complexity Theory for Usable Security. In *Proceedings of the 2015 New Security Paradigms Workshop (NSPW '15)*. Association for Computing Machinery, New York, NY, USA, 85–99. https://doi.org/10.1145/2841113.2841120

[16] Mark C. Bolino and William H. Turnley. 1999. Measuring Impression Management in Organizations: A Scale Development Based on the Jones and Pittman Taxonomy. *Organizational Research Methods* 2, 2 (April 1999), 187–206. https://doi.org/10.1177/109442819922005 Publisher: SAGE Publications Inc.

[17] Grant M. Brady, Donald M. Truxillo, Talya N. Bauer, and Mark P. Jones. 2021. The development and validation of the Privacy and Data Security Concerns Scale (PDSCS). *International Journal of Selection and Assessment* 29, 1 (2021), 100–113. https://doi.org/10.1111/ijsa.12311 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ijsa.12311.

[18] John Brook. 1996. SUS: a" quick and dirty" usability scale. *Usability evaluation in industry* (1996). Publisher: Taylor and Francis.

[19] Simone Browne. 2015. *Dark Matters: On the Surveillance of Blackness*. Duke University Press. Google-Books-ID: snmJCgAAQBAJ.

[20] Finn Brunton and Helen Fay Nissenbaum. 2015. *Obfuscation: a user's guide for privacy and protest*. MIT Press, Cambridge, Massachusetts.

[21] James Carifio and Rocco Perla. 2008. Resolving the 50-year debate around using and misusing Likert scales. *Medical Education* 42, 12 (Dec. 2008), 1150–1152. https://doi.org/10.1111/j.1365-2923.2008.03172.x

[22] Kevin D. Carlson and Andrew O. Herdman. 2012. Understanding the Impact of Convergent Validity on Research Results. *Organizational Research Methods* 15, 1 (Jan. 2012), 17–32. https://doi.org/10.1177/1094428110392383

[23] Stevie Chancellor, Eric P. S. Baumer, and Munmun De Choudhury. 2019. Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 147:1–147:32. https://doi.org/10.1145/3359249

[24] Varun Chandrasekaran, Chuhan Gao, Brian Tang, Kassem Fawaz, Somesh Jha, and Suman Banerjee. 2021. Face-Off: Adversarial Face Obfuscation. *Proceedings on Privacy Enhancing Technologies* 2021, 2 (2021), 369–390.

[25] Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John Dickerson, Gavin Taylor, and Tom Goldstein. 2021. LowKey: Leveraging Adversarial Attacks to Protect Social Media Users from Facial Recognition. *arXiv:2101.07922 [cs]* (Jan. 2021). http://arxiv.org/abs/2101.07922 arXiv: 2101.07922.

[26] Alexander Cho. 2018. Default publicness: Queer youth of color, social media, and being outed by the machine. *New Media & Society* 20, 9 (Sept. 2018), 3183–3200. https://doi.org/10.1177/1461444817744784 Publisher: SAGE Publications.

[27] Chhaya Chouhan, Christy M. LaPerriere, Zaina Aljallad, Jess Kropczynski, Heather Lipford, and Pamela J. Wisniewski. 2019. Co-designing for Community Oversight: Helping People Make Privacy and Security Decisions Together. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 146:1–146:31. https://doi.org/10.1145/3359248

[28] Mitchell Church, Ravi Thambusamy, and Hamid Nemati. 2020. User misrepresentation in online social networks: how competition and altruism impact online disclosure behaviours. *Behaviour & Information Technology* 39, 12 (Dec. 2020), 1320–1340. https://doi.org/10.1080/0144929X.2019.1667440 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0144929X.2019.1667440.

[29] Kovila P.L. Coopamootoo. 2020. Usage Patterns of Privacy-Enhancing Technologies. In *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security (CCS '20)*. Association for Computing Machinery, New York, NY, USA, 1371–1390. https://doi.org/10.1145/3372297.3423347

[30] E. D. Cristofaro, C. Soriente, G. Tsudik, and A. Williams. 2012. Hummingbird: Privacy at the Time of Twitter. In *2012 IEEE Symposium on Security and Privacy*. 285–299. https://doi.org/10.1109/SP.2012.26 ISSN: 2375-1207.

[31] Sauvik Das. 2020. Subversive AI: Resisting automated algorithmic surveillance with human-centered adversarial machine learning. In *Resistance AI workshop at NeurIPS*. 4.

[32] Sauvik Das, Cori Faklaris, Jason I Hong, Laura A Dabbish, and others. 2022. The security & privacy acceptance framework (spaf). *Foundations and Trends® in Privacy and Security* 5, 1-2 (2022), 1–143. Publisher: Now Publishers, Inc..

[33] Sauvik Das, Tiffany Hyun-Jin Kim, Laura A. Dabbish, and Jason I. Hong. 2014. The Effect of Social Influence on Security Sensitivity. 143–157. https://www.usenix.org/conference/soups2014/proceedings/presentation/das

[34] Sauvik Das, Adam D.I. Kramer, Laura A. Dabbish, and Jason I. Hong. 2015. The Role of Social Influence in Security Feature Adoption. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. Association for Computing Machinery, Vancouver, BC, Canada, 1416–1426. https://doi.org/10.1145/2675133.2675225

[35] Julia Davies. 2007. Display, Identity and the Everyday: Self-presentation through online image sharing. *Discourse: studies in the cultural politics of education* 28, 4 (2007), 549–564. Publisher: Taylor & Francis.

[36] Fred D Davis. 1985. *A technology acceptance model for empirically testing new end-user information systems: Theory and results*. phd. Massachusetts Institute of Technology.

[37] Robert F DeVellis and Carolyn T Thorpe. 2021. *Scale development: Theory and applications*. Sage publications.

[38] Amandeep Dhir. 2017. Why Do Young People Avoid Photo Tagging? A New Service Avoidance Scale. *Social Science Computer Review* 35, 4 (Aug. 2017), 480–497. https://doi.org/10.1177/0894439316653636 Publisher: SAGE Publications Inc.

[39] Angel Diaz. 2019. *New York City Police Department Surveillance Technology.* Technical Report. Brennan Center for Justice. https://www.brennancenter.org/our-work/research-reports/new-york-city-police-department-surveillance-technology

[40] Youngwook Do, Nivedita Arora, Ali Mirzazadeh, Injoo Moon, Eryue Xu, Zhihan Zhang, Gregory D Abowd, and Sauvik Das. 2023. Powering for privacy: Improving user trust in smart speaker microphones with intentional powering and perceptible assurance. (2023).

[41] Youngwook Do, Jung Wook Park, Yuxi Wu, Avinandan Basu, Dingtian Zhang, Gregory D Abowd, and Sauvik Das. 2021. Smart Webcam Cover: Exploring the Design of an Intelligent Webcam Cover to Improve Usability and Trust. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 4 (2021), 1–21. Publisher: ACM New York, NY, USA.

[42] Serge Egelman and Eyal Peer. 2015. Scaling the Security Wall: Developing a Security Behavior Intentions Scale (SeBIS). In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 2873–2882. https://doi.org/10.1145/2702123.2702249

[43] Upol Ehsan and Mark O. Riedl. 2020. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence*, Constantine Stephanidis, Masaaki Kurosu, Helmut Degen, and Lauren Reinerman-Jones (Eds.). Springer International Publishing, Cham, 449–466. https://doi.org/10.1007/978-3-030-60117-1_33

[44] Cori Faklaris, Laura A Dabbish, and Jason I Hong. 2019. A {self-report} measure of {end-user} security attitudes ({{{{{SA-6}}}}}). In *Fifteenth symposium on usable privacy and security (SOUPS 2019)*. 61–77.

[45] Jesse Fox and Megan A. Vendemia. 2016. Selective Self-Presentation and Social Comparison Through Photographs on Social Networking Sites. *Cyberpsychology, Behavior, and Social Networking* 19, 10 (Oct. 2016), 593–600. https://doi.org/10.1089/cyber.2016.0248 Publisher: Mary Ann Liebert, Inc., publishers.

[46] S. M. Furnell, D. Katsabas, P. S. Dowland, and F. Reid. 2007. A Practical Usability Evaluation of Security Features in End-User Applications. In *New Approaches for Security, Privacy and Trust in Complex Environments (IFIP International Federation for Information Processing)*, Hein Venter, Mariki Eloff, Les Labuschagne, Jan Eloff, and Rossouw von Solms (Eds.). Springer US, Boston, MA, 205–216. https://doi.org/10.1007/978-0-387-72367-9_18

[47] Vaibhav Garg, Kevin Benton, and L. Jean Camp. 2014. *The Privacy Paradox: A Facebook Case Study.* SSRN Scholarly Paper ID 2411672. Social Science Research Network, Rochester, NY. https://doi.org/10.2139/ssrn.2411672

[48] Nina Gerber, Verena Zimmermann, Birgit Henhapl, Sinem Emeröz, and Melanie Volkamer. 2018. Finally Johnny Can Encrypt: But Does This Make Him Feel More Secure?. In *Proceedings of the 13th International Conference on Availability, Reliability and Security (ARES 2018)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3230833.3230859

[49] Saikat Guha, Kevin Tang, and Paul Francis. 2008. NOYB: privacy in online social networks. In *Proceedings of the first workshop on Online social networks (WOSN '08)*. Association for Computing Machinery, New York, NY, USA, 49–54. https://doi.org/10.1145/1397735.1397747

[50] Jan Philip Göpfert, André Artelt, Heiko Wersing, and Barbara Hammer. 2020. Adversarial Attacks Hidden in Plain Sight. In *Advances in Intelligent Data Analysis XVIII (Lecture Notes in Computer Science)*, Michael R. Berthold, Ad Feelders, and Georg Krempl (Eds.). Springer International Publishing, Cham, 235–247. https://doi.org/10.1007/978-3-030-44584-3_19

[51] Carol Hall and Amanda Roshier. 2016. Getting the measure of behavior… is seeing believing? *Interactions* 23, 4 (2016), 42–46. Publisher: ACM New York, NY, USA.

[52] Cory Hallam and Gianluca Zanella. 2017. Online self-disclosure: The privacy paradox explained as a temporally discounted balance between concerns and rewards. *Computers in Human Behavior* 68 (March 2017), 217–227. https://doi.org/10.1016/j.chb.2016.11.033

[53] Mitchell M. Handelsman, William L. Briggs, Nora Sullivan, and Annette Towler. 2005. A Measure of College Student Course Engagement. *The Journal of Educational Research* 98, 3 (Jan. 2005), 184–192. https://doi.org/10.3200/JOER.98.3.184-192 Publisher: Routledge _eprint: https://doi.org/10.3200/JOER.98.3.184-192.

[54] Rakibul Hasan, Bennett I. Bertenthal, Kurt Hugenberg, and Apu Kapadia. 2021. Your Photo is so Funny that I don't Mind Violating Your Privacy by Sharing it: Effects of Individual Humor Styles on Online Photo-sharing Behaviors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–14. https://doi.org/10.1145/3411764.3445258

[55] Rakibul Hasan, Rudolf Siegel, Rebecca Weil, and Katharina Krombholz. 2022. Developing a psychometric scale to measure one's valuation of other people's privacy. In *The eighteenth symposium on usable privacy and security (SOUPS 2022)*.

[56] Hassan, Rakibul Hasan, Patrick Shaffer, David Crandall, and Eman T. Apu Kapadia. 2017. Cartooning for Enhanced Privacy in Lifelogging and Streaming Videos. 29–38. https://openaccess.thecvf.com/content_cvpr_2017_workshops/w16/html/Kapadia_Cartooning_for_Enhanced_CVPR_2017_paper.html

[57] J. He, B. Liu, D. Kong, X. Bao, N. Wang, H. Jin, and G. Kesidis. 2016. PUPPIES: Transformation-Supported Personalized Privacy Preserving Partial Image Sharing. In *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. 359–370. https://doi.org/10.1109/DSN.2016.40 ISSN: 2158-3927.

[58] Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. Most people are not WEIRD. *Nature* 466, 7302 (2010), 29–29. Publisher: Nature Publishing Group.

[59] Joseph Henrich, Steven J. Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and Brain Sciences* 33, 2-3 (June 2010), 61–83. https://doi.org/10.1017/S0140525X0999152X

[60] Amir Herzberg and Hemi Leibowitz. 2016. Can Johnny finally encrypt? evaluating E2E-encryption in popular IM applications. In *Proceedings of the 6th Workshop on Socio-Technical Aspects in Security and Trust (STAST '16)*. Association for Computing Machinery, New York, NY, USA, 17–28. https://doi.org/10.1145/3046055.3046059

[61] Kashmir Hill. 2020. This Tool Could Protect Your Photos From Facial Recognition. *The New York Times* (Aug. 2020). https://www.nytimes.com/2020/08/03/technology/fawkes-tool-protects-photos-from-facial-recognition.html

[62] Timothy R. Hinkin. 1998. A Brief Tutorial on the Development of Measures for Use in Survey Questionnaires. *Organizational Research Methods* 1, 1 (Jan. 1998), 104–121. https://doi.org/10.1177/109442819800100106

[63] Chin-Chang Ho and Karl F. MacDorman. 2017. Measuring the Uncanny Valley Effect: Refinements to Indices for Perceived Humanness, Attractiveness, and Eeriness. *International Journal of Social Robotics* 9, 1 (Jan. 2017), 129–139. https://doi.org/10.1007/s12369-016-0380-9

[64] Daniel T. Holt, Achilles A. Armenakis, Hubert S. Feild, and Stanley G. Harris. 2007. Readiness for Organizational Change: The Systematic Development of a Scale. *The Journal of Applied Behavioral Science* 43, 2 (June 2007), 232–255. https://doi.org/10.1177/0021886306295295 Publisher: SAGE Publications Inc.

[65] Andreas Holzinger, Gig Searle, Thomas Kleinberger, Ahmed Seffah, and Homa Javahery. 2008. Investigating Usability Metrics for the Design and Development of Applications for the Elderly. In *Computers Helping People with Special Needs (Lecture Notes in Computer Science)*, Klaus Miesenberger, Joachim Klaus, Wolfgang Zagler, and Arthur Karshmer (Eds.). Springer, Berlin, Heidelberg, 98–105. https://doi.org/10.1007/978-3-540-70540-6_13

[66] Seoyeon Hong, Mi R. Jahng, Namyeon Lee, and Kevin R. Wise. 2020. Do you filter who you are?: Excessive self-presentation, social cues, and user evaluations of Instagram selfies. *Computers in Human Behavior* 104 (March 2020), 106159. https://doi.org/10.1016/j.chb.2019.106159

[67] Roberto Hoyle, Luke Stark, Qatrunnada Ismail, David Crandall, Apu Kapadia, and Denise Anthony. 2020. Privacy Norms and Preferences for Photos Posted Online. *ACM Transactions on Computer-Human Interaction* 27, 4 (Sept. 2020), 1–27. https://doi.org/10.1145/3380960

[68] Yuheng Hu, Lydia Manikonda, and Subbarao Kambhampati. 2014. What We Instagram: A First Analysis of Instagram Photo Content and User Types. *Proceedings of the International AAAI Conference on Web and Social Media* 8, 1 (May 2014), 595–598. https://doi.org/10.1609/icwsm.v8i1.14578 Number: 1.

[69] Azham Hussain and Maria Kutar. 2009. Usability Metric Framework for Mobile Phone Application. *PGNet, ISBN* 2099 (2009), 978–1.

[70] Sergio Ibáñez-Sánchez, Carlos Orús, and Carlos Flavián. 2022. Augmented reality filters on social media. Analyzing the drivers of playability based on uses and gratifications theory. *Psychology & Marketing* 39, 3 (2022), 559–578. https://doi.org/10.1002/mar.21639 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mar.21639.

[71] Indu Ilanchezian, Praneeth Vepakomma, Abhishek Singh, Otkrist Gupta, G. N. Srinivasa Prasanna, and Ramesh Raskar. 2019. Maximal adversarial perturbations for obfuscation: Hiding certain attributes while preserving rest. https://doi.org/10.48550/arXiv.1909.12734 arXiv:1909.12734 [cs, stat].

[72] 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 (March 2022), 107126. https://doi.org/10.1016/j.chb.2021.107126

[73] Ronald Kainda, Ivan Fléchais, and A.W. Roscoe. 2010. Security and Usability: Analysis and Evaluation. In *2010 International Conference on Availability, Reliability and Security*. 275–282. https://doi.org/10.1109/ARES.2010.77

[74] Ruogu Kang, Stephanie Brown, Laura Dabbish, and Sara Kiesler. 2014. Privacy Attitudes of Mechanical Turk Workers and the U.S. Public. 37–49. https://www.usenix.org/conference/soups2014/proceedings/presentation/kang

[75] Naz Kaya and Margaret J. Weber. 2003. Cross-cultural differences in the perception of crowding and privacy regulation: American and Turkish students. *Journal of Environmental Psychology* 23, 3 (Sept. 2003), 301–309. https://doi.org/10.1016/S0272-4944(02)00087-7

[76] Brian Keelan. 2002. *Handbook of image quality: characterization and prediction*. CRC Press.

[77] Katharina Krombholz, Adrian Dabrowski, Matthew Smith, and Edgar Weippl. 2015. Ok Glass, Leave Me Alone: Towards a Systematization of Privacy Enhancing Technologies for Wearable Computing. In *Financial Cryptography and Data Security (Lecture Notes in Computer Science)*, Michael Brenner, Nicolas Christin, Benjamin Johnson, and Kurt Rohloff (Eds.). Springer, Berlin, Heidelberg, 274–280. https://doi.org/10.1007/978-3-662-48051-9_20

[78] H. A. Kruger, L. Drevin, S. Flowerday, and T. Steyn. 2011. An assessment of the role of cultural factors in information security awareness. In *2011 Information Security for South Africa*. 1–7. https://doi.org/10.1109/ISSA.2011.6027505 ISSN: 2330-9881.

[79] Yao Li, Eugenia Ha Rim Rho, and Alfred Kobsa. 2022. Cultural differences in the effects of contextual factors and privacy concerns on users' privacy decision on social networking sites. *Behaviour & Information Technology* 41, 3 (Feb. 2022), 655–677. https://doi.org/10.1080/0144929X.2020.1831608 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0144929X.2020.1831608.

[80] 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 (Dec. 2017), 67:1–67:24. https://doi.org/10.1145/3134702

[81] Q. Vera Liao and Kush R. Varshney. 2022. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. *arXiv:2110.10790 [cs]* (Jan. 2022). http://arxiv.org/abs/2110.10790 arXiv: 2110.10790.

[82] Jiang Liu, Chun Pong Lau, and Rama Chellappa. 2023. DiffProtect: Generate Adversarial Examples with Diffusion Models for Facial Privacy Protection. _eprint: 2305.13625.

[83] Liu Liu, Olivier De Vel, Qing-Long Han, Jun Zhang, and Yang Xiang. 2018. Detecting and preventing cyber insider threats: A survey. *IEEE Communications Surveys & Tutorials* 20, 2 (2018), 1397–1417. Publisher: IEEE.

[84] Zhuoran Liu, Zhengyu Zhao, and Martha A Larson. 2019. Pixel Privacy 2019: Protecting Sensitive Scene Information in Images.. In *MediaEval*.

[85] Jacob Logas, Ari Schlesinger, Zhouyu Li, and Sauvik Das. 2022. Image DePO: towards gradual decentralization of online social networks using decentralized privacy overlays. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–28. Publisher: ACM New York, NY, USA.

[86] Ryan Mac, Caroline Haskins, and Logan McDonald. 2020. Clearview's Facial Recognition App Has Been Used By The Justice Department, ICE, Macy's, Walmart, And The NBA. https://www.buzzfeednews.com/article/ryanmac/clearview-ai-fbi-ice-global-law-enforcement

[87] Penousal Machado and Amilcar Cardoso. 1998. Computing Aesthetics. In *Advances in Artificial Intelligence (Lecture Notes in Computer Science)*, Flávio Moreira de Oliveira (Ed.). Springer, Berlin, Heidelberg, 219–228. https://doi.org/10.1007/10692710_23

[88] Naresh K. Malhotra, Sung S. Kim, and James Agarwal. 2004. Internet Users' Information Privacy Concerns (IUIPC): The Construct, the Scale, and a Causal Model. *Information Systems Research* 15, 4 (Dec. 2004), 336–355. https://doi.org/10.1287/isre.1040.0032 Publisher: INFORMS.

[89] Nur MERDANOĞLU and Pınar ONAY DURDU. 2018. A systematic mapping study of usability vs security. In *2018 6th International Conference on Control Engineering & Information Technology (CEIT)*. 1–6. https://doi.org/10.1109/CEIT.2018.8751841

[90] Jacqueline Murray. 2013. Likert Data: What to use, parametric or non-parametric? 4, 11 (2013).

[91] Moses Namara, Daricia Wilkinson, Kelly Caine, and Bart P. Knijnenburg. 2020. Emotional and Practical Considerations Towards the Adoption and Abandonment of VPNs as a Privacy-Enhancing Technology. *Proceedings on Privacy Enhancing Technologies* 2020, 1 (Jan. 2020), 83–102. https://doi.org/10.2478/popets-2020-0006

[92] Geoffrey Norman. 2010. Likert scales, levels of measurement and the "laws" of statistics. *Advances in health sciences education : theory and practice* 15 (Feb. 2010), 625–32. https://doi.org/10.1007/s10459-010-9222-y

[93] Melvin R Novick. 1966. The axioms and principal results of classical test theory. *Journal of mathematical psychology* 3, 1 (1966), 1–18. Publisher: Elsevier.

[94] Anne Oeldorf-Hirsch and S. Shyam Sundar. 2016. Social and Technological Motivations for Online Photo Sharing. *Journal of Broadcasting & Electronic Media* 60, 4 (Oct. 2016), 624–642. https://doi.org/10.1080/08838151.2016.1234478 Publisher: Routledge _eprint: https://doi.org/10.1080/08838151.2016.1234478.

[95] Michael A. Oren and Stephen B. Gilbert. 2011. Framework for measuring social affinity for CSCW software. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems (CHI EA '11)*. Association for Computing Machinery, New York, NY, USA, 1387–1392. https://doi.org/10.1145/1979742.1979779

[96] Hilarie Orman. 2015. Why Won't Johnny Encrypt? *IEEE Internet Computing* 19, 1 (Jan. 2015), 90–94. https://doi.org/10.1109/MIC.2015.16 Conference Name: IEEE Internet Computing.

[97] Greg Orr. 2003. Diffusion of innovations, by Everett Rogers (1995). *Retrieved January* 21 (2003), 2005.

[98] Xinru Page, Reza Ghaiumy Anaraky, Bart P. Knijnenburg, and Pamela J. Wisniewski. 2019. Pragmatic Tool vs. Relational Hindrance: Exploring Why Some Social Media Users Avoid Privacy Features. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 110:1–110:23. https://doi.org/10.1145/3359212

[99] Yilang Peng. 2017. Time Travel with One Click: Effects of Digital Filters on Perceptions of Photographs. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 6000–6011. https://doi.org/10.1145/3025453.3025810

[100] 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. Publisher: ACM New York, NY, USA.

[101] Jeffrey G. Proudfoot, David Wilson, Joseph S. Valacich, and Michael D. Byrd. 2018. Saving face on Facebook: privacy concerns, social benefits, and impression management. *Behaviour & Information Technology* 37, 1 (Jan. 2018), 16–37. https://doi.org/10.1080/0144929X.2017.1389988 Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0144929X.2017.1389988.

[102] Moo-Ryong Ra, Ramesh Govindan, and Antonio Ortega. 2013. P3: Toward Privacy-Preserving Photo Sharing. 515–528. https://www.usenix.org/conference/nsdi13/technical-sessions/presentation/ra

[103] Emilee Rader, Rick Wash, and Brandon Brooks. 2012. Stories as informal lessons about security. In *Proceedings of the Eighth Symposium on Usable Privacy and Security (SOUPS '12)*. Association for Computing Machinery, Washington, D.C., 1–17. https://doi.org/10.1145/2335356.2335364

[104] Evani Radiya-Dixit and Florian Tramèr. 2021. Data Poisoning Won't Save You From Facial Recognition. *arXiv:2106.14851 [cs]* (June 2021). http://arxiv.org/abs/2106.14851 arXiv: 2106.14851.

[105] Arezoo Rajabi, Rakesh B Bobba, Mike Rosulek, Charles Wright, and Wu-chi Feng. 2021. On the (im) practicality of adversarial perturbation for image privacy. *Proceedings on Privacy Enhancing Technologies* (2021).

[106] Christian Remy, Oliver Bates, Alan Dix, Vanessa Thomas, Mike Hazas, Adrian Friday, and Elaine M. Huang. 2018. Evaluation Beyond Usability: Validating Sustainable HCI Research. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173790

[107] Sebastian Ruder. 2017. An overview of gradient descent optimization algorithms. arXiv: 1609.04747 [cs.LG].

[108] Jian Raymond Rui and Michael A Stefanone. 2013. Strategic image management online: Self-presentation, self-esteem and social network perspectives. *Information, Communication & Society* 16, 8 (2013), 1286–1305. Publisher: Taylor & Francis.

[109] Scott Ruoti, Jeff Andersen, Daniel Zappala, and Kent Seamons. 2016. Why Johnny Still, Still Can't Encrypt: Evaluating the Usability of a Modern PGP Client. (Jan. 2016). http://arxiv.org/abs/1510.08555 arXiv:1510.08555 [cs].

[110] Johnny Saldaña. 2009. *The coding manual for qualitative researchers*. Sage, Los Angeles, Calif. OCLC: ocn233937452.

[111] Ilyssa Salomon and Christia Spears Brown. 2021. That selfie becomes you: examining taking and posting selfies as forms of self-objectification. *Media Psychology* 24, 6 (Nov. 2021), 847–865. https://doi.org/10.1080/15213269.2020.1817091 Publisher: Routledge _eprint: https://doi.org/10.1080/15213269.2020.1817091.

[112] Ayon Sen, Xiaojin Zhu, Erin Marshall, and Robert Nowak. 2020. Popular Imperceptibility Measures in Visual Adversarial Attacks are Far from Human Perception. In *Decision and Game Theory for Security (Lecture Notes in Computer Science)*, Quanyan Zhu, John S. Baras, Radha Poovendran, and Juntao Chen (Eds.). Springer International Publishing, Cham, 188–199. https://doi.org/10.1007/978-3-030-64793-3_10

[113] Ayon Sen, Xiaojin Zhu, Liam Marshall, and Robert Nowak. 2019. Should Adversarial Attacks Use Pixel p-Norm? https://doi.org/10.48550/arXiv.1906.02439 arXiv:1906.02439 [cs, stat].

[114] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y. Zhao. 2023. Glaze: Protecting Artists from Style Mimicry by Text-to-Image Models. _eprint: 2302.04222.

[115] Shawn Shan, Emily Wenger, Jiayun Zhang, Huiying Li, Haitao Zheng, and Ben Y Zhao. 2020. Fawkes: Protecting Privacy against Unauthorized Deep Learning Models. 16.

[116] Mahmood Sharif, Lujo Bauer, and Michael K. Reiter. 2018. On the Suitability of Lp-Norms for Creating and Preventing Adversarial Examples. 1605–1613. https://openaccess.thecvf.com/content_cvpr_2018_workshops/w32/html/Sharif_On_the_Suitability_CVPR_2018_paper.html

[117] Ernestasia Siahaan, Alan Hanjalic, and Judith Redi. 2016. A Reliable Methodology to Collect Ground Truth Data of Image Aesthetic Appeal. *IEEE Transactions on Multimedia* 18, 7 (July 2016), 1338–1350. https://doi.org/10.1109/TMM.2016.2559942 Conference Name: IEEE Transactions on Multimedia.

[118] Stephen G. Sireci. 1998. The Construct of Content Validity. *Social Indicators Research* 45, 1/3 (1998), 83–117. https://www.jstor.org/stable/27522338 Publisher: Springer.

[119] Manya Sleeper, Rebecca Balebako, Sauvik Das, Amber Lynn McConahy, Jason Wiese, and Lorrie Faith Cranor. 2013. The post that wasn't: exploring self-censorship on facebook. In *Proceedings of the 2013 conference on Computer supported cooperative work (CSCW '13)*. Association for Computing Machinery, New York, NY, USA, 793–802. https://doi.org/10.1145/2441776.2441865

[120] Elizabeth Stoycheff, G. Scott Burgess, and Maria Clara Martucci. 2020. Online censorship and digital surveillance: the relationship between suppression technologies and democratization across countries. *Information, Communication & Society* 23, 4 (March 2020), 474–490. https://doi.org/10.1080/1369118X.2018.1518472 Publisher: Routledge _eprint: https://doi.org/10.1080/1369118X.2018.1518472.

[121] Priyanka Surendran. 2012. Technology Acceptance Model: A Survey of Literature. *International Journal of Business and Social Research* (2012).

[122] Bernadette Szajna. 1996. Empirical Evaluation of the Revised Technology Acceptance Model. *Management Science* 42, 1 (Jan. 1996), 85–92. https://doi.org/10.1287/mnsc.42.1.85

[123] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. http://arxiv.org/abs/1312.6199 arXiv:1312.6199 [cs].

[124] Monika Taddicken. 2014. The 'Privacy Paradox' in the Social Web: The Impact of Privacy Concerns, Individual Characteristics, and the Perceived Social Relevance on Different Forms of Self-Disclosure*. *Journal of Computer-Mediated Communication* 19, 2 (Jan. 2014), 248–273. https://doi.org/10.1111/jcc4.12052

[125] Miguel A. Teruel, Elena Navarro, Víctor López-Jaquero, Francisco Montero, and Pascual González. 2014. A CSCW Requirements Engineering CASE Tool: Development and usability evaluation. *Information and Software Technology* 56, 8 (Aug. 2014), 922–949. https://doi.org/10.1016/j.infsof.2014.02.009

[126] Kim-Han Thung and Paramesran Raveendran. 2009. A survey of image quality measures. In *2009 International Conference for Technical Postgraduates (TECHPOS)*. 1–4. https://doi.org/10.1109/TECHPOS.2009.5412098

[127] Alise Tifentale and Lev Manovich. 2015. Selfiecity: Exploring Photography and Self-Fashioning in Social Media. In *Postdigital Aesthetics: Art, Computation and Design*, David M. Berry and Michael Dieter (Eds.). Palgrave Macmillan UK, London, 109–122. https://doi.org/10.1057/9781137437204_9

[128] Ledyard R Tucker and Robert C MacCallum. 1997. Exploratory factor analysis. *Unpublished manuscript, Ohio State University, Columbus* (1997), 1–459.

[129] Megan A. Vendemia and David C. DeAndrea. 2021. The effects of engaging in digital photo modifications and receiving favorable comments on women's selfies shared on social media. *Body Image* 37 (June 2021), 74–83. https://doi.org/10.1016/j.bodyim.2021.01.011

[130] Serena Wang, Cori Faklaris, Junchao Lin, Laura Dabbish, and Jason I Hong. 2022. 'It's problematic but I'm not concerned': University perspectives on account sharing. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–27. Publisher: ACM New York, NY, USA.

[131] Emily Wenger, Shawn Shan, Haitao Zheng, and Ben Y. Zhao. 2023. SoK: Anti-Facial Recognition Technology. _eprint: 2112.04558.

[132] Alma Whitten and J Doug Tygar. 1999. Why Johnny Can't Encrypt: A Usability Evaluation of PGP 5.0.. In *USENIX Security Symposium*, Vol. 348. 169–184.

[133] Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu, Yuefeng Chen, and Hui Xue. 2021. Towards Face Encryption by Generating Adversarial Identity Masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 3897–3907.

[134] Svetlana Yarosh, Panos Markopoulos, and Gregory D. Abowd. 2014. Towards a questionnaire for measuring affective benefits and costs of communication technologies. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing (CSCW '14)*. Association for Computing Machinery, New York, NY, USA, 84–96. https://doi.org/10.1145/2531602.2531634

[135] J. Yu, Z. Kuang, Z. Yu, D. Lin, and J. Fan. 2017. Privacy Setting Recommendation for Image Sharing. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 726–730. https://doi.org/10.1109/ICMLA.2017.00-73

[136] J. Yu, B. Zhang, Z. Kuang, D. Lin, and J. Fan. 2017. iPrivacy: Image Privacy Protection by Identifying Sensitive Objects via Deep Multi-Task Learning. *IEEE Transactions on Information Forensics and Security* 12, 5 (May 2017), 1005–1016. https://doi.org/10.1109/TIFS.2016.2636090 Conference Name: IEEE Transactions on Information Forensics and Security.

[137] Abbas Zabihzadeh, Mohammad Ali Mazaheri, Javad Hatami, Mohammad Reza Nikfarjam, Leili Panaghi, and Telli Davoodi. 2019. Cultural differences in conceptual representation of "Privacy": A comparison between Iran and the United States. *The Journal of Social Psychology* 159, 4 (July 2019), 357–370. https://doi.org/10.1080/00224545.2018.1493676 Publisher: Routledge _eprint: https://doi.org/10.1080/00224545.2018.1493676.

[138] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. 2019. Dissonance Between Human and Machine Understanding. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (Nov. 2019), 56:1–56:23. https://doi.org/10.1145/3359158

[139] Dorothy Zhao, Mikako Inaba, and Andrés Monroy-Hernández. 2022. Understanding teenage perceptions and configurations of privacy on instagram. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2 (Nov. 2022). https://doi.org/10.1145/3555608 Number of pages: 28 Place: New York, NY, USA Publisher: Association for Computing Machinery tex.articleno: 550 tex.issue_date: November 2022.

[140] Shoshana Zuboff. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs. Google-Books-ID: lRqrDQAAQBAJ.

[141] Umur A. Çiftçi, Gokturk Yuksek, and İlke Demir. 2023. My Face My Choice: Privacy Enhancing Deepfakes for Social Media Anonymization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 1369–1379.

A SAIA-8: RESEARCH CONSTRUCTS AND MEASURES

A.1 Image Aesthetic

Five-point scale anchored with "strongly disagree" and "strongly agree."

(1) I don't feel comfortable with the changes made to the photograph.
(2) I feel concerned with how the filter has affected my looks.
(3) I feel the filtered photo's changes are immediately noticable.
(4) My family or friends would ask about the filtered photo if I posted it on social media.

A.2 Identity Modification

Five-point scale anchored with "strongly disagree" and "strongly agree."

(1) The filter makes me look less human.

A.3 Willingness to Share

Five-point scale anchored with "strongly disagree" and "strongly agree."

(1) The changes made by the filter defeat the purpose of sharing the image.
(2) I wouldn't share the image publicly after the filter was applied.
(3) I would rarely use this filter for photos shared to social media.

B QUALITATIVE EXPLORATION STUDY

B.1 Part 1: Orienting

(1) Do you use any privacy preserving tools at all? (e.g., Incognito mode, VPN, Private accounts, Secure messaging, TOR browser, Ad blocker) **Note**: if multiple given choose the one with the most institutional protection.
(2) Can you recall a specific time when you have used X?
(3) Why did you use X in this instance?
(4) Was there any event or incident that encouraged you to use X?
(5) With the photo you chose, why were you concerned about sharing that photo online?
(6) If you had to post the photo you chose, how would you protect yourself?

B.2 Part 2: Perturbation

In this part of the interview, we will introduce you to tools that preserve your privacy by making changes to the faces in a photo. These changes make it such that automated systems cannot recognize the subject. The tools aim to make it more difficult for untrusted third parties to search for and view personal photos posted online. Think of it as similar to the tools you already use (e.g., incognito mode, VPN, private accounts) to keep your information private.

(1) Do you notice any changes to the photo? If so, can you describe them?
(2) How do you feel about the changes made to your photo?
(3) How do the changes make you feel about sharing on social media?
(4) Does the promise of protection afforded by the changes made at all impact your willingness to share the photo on social media?
(5) (In the case of multiple qualities) Assuming higher settings provide more privacy protection, which would you be willing to choose and why?

(6) Think of the last image you did share on social media, how would you feel sharing it after putting it through this tool?
(7) Do you have any questions or concerns about this tool and its output?

Repeat (Standard Obfuscation, Fawkes, LowKey)

B.3 Part 3: Preference

(1) Which output is your favorite of the ones you have seen? What is your least favorite? Why?
(2) What are differences that immediately come to mind between the models?
(3) Given this image is sensitive, how would you make a decision on which one to share?
(4) What would need to change for you to prefer model (X) over model (Y)?
(5) Are there different situations you could see yourself choosing one method over the other?

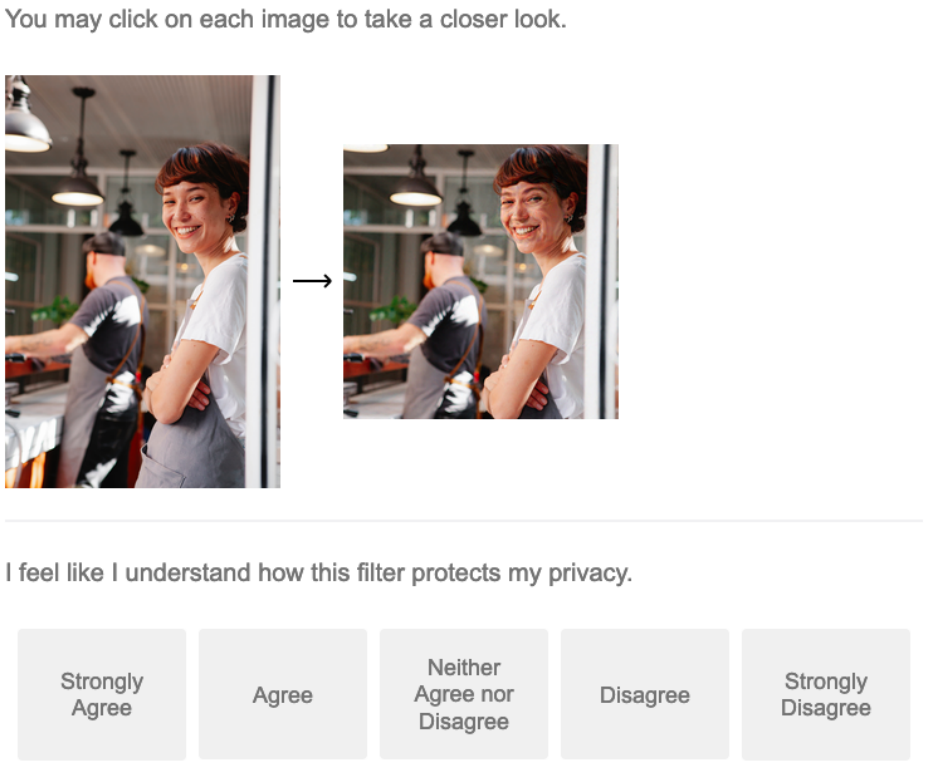
C CANDIDATE ITEMS

ID	Question
1	I feel like the filtered image does not look like me.
2	The filtered image looks like me.
3	The filtered image reflects how I see myself.
4	The filtered image reflects how I want to be seen.
5	I feel like the filtered image protects my identity.
6	I believe the filtered image obscures my identity.
7	People who know me can tell that I am in the filtered image.
8	The image has been modified in a positive way.
9	The image has been modified in a negative way.
10	I prefer the filtered image over the original.
11	The filtered photo has a noticeable difference from the original.
12	I can see myself using this filter whenever I want to protect my privacy.
13	My face is the same as it is in the original image.
14	I would like to use this filter even if it didn't protect my privacy.
15	I feel skeptical of the amount of privacy protection this filter provides.
16	My family or friends would ask about the filtered photo if I posted it on social media.
17	My friends would ask about the filtered photo if I posted it on social media.
18	This filter reflects my preferred gender presentation.
19	This filter has feminized my face.
20	This filter has made my face look more masculine.
21	The filter reflects my natural skin tone.
22	The filter changes my skin tone in a pleasing way.
23	The changes made by the filter are coherent with the rest of the image.
24	I feel like I am recognizable in the filtered image.
25	I don't recognize myself in the filtered image.
26	The filtered image is an authentic representation of me.
27	The changes made by the filter defeat the purpose of sharing the image.
28	I wouldn't share the image publicly after the filter was applied.
29	I look more mature after the filter is applied.
30	I look older after the filter was applied.
31	I look younger after the filter was applied.
32	I look more healthy after the filter was applied.
33	I look less healthy after the filter was applied.

34	The filter draws attention to specific areas of the image.
35	The filter retains the humanity of the photo's subject.
36	The filter objectifies the photo subject.
37	I look closer to my ideal representation over the original.
38	The filter is affirming to my gender identity.
39	I am confident that this filter will protect my identity.
40	This filter matches my expectation for a privacy preserving tool.
41	The filtered image will protect me against recognition by third parties.
42	I am satisfied by the filtered image.
43	I would have to explain my appearance if I shared this photo on social media.
44	I can see myself using this filter on all my photos.
45	I would be happy if this filter was automatically applied to all my pictures.
46	The changes made by this filter are surprising.
47	The changes made by this filter match the rest of the photo.
48	I feel the filtered photo's changes are immediately noticeable.
49	The filter worsens the color in the photo.
50	I feel like I would need to edit the filtered photo more.
51	I would post the filtered photo on social media after photoshopping it.
52	I don't like the filtered photo but the privacy it provides gives me confidence to post it online.
53	I am uncomfortable with the filtered photo, but the increased privacy gives me more comfort to post it online.
54	I feel comfortable with publicly posting the filtered photo online, if it improves my privacy.
55	I am unwilling to publicly post the filtered photo online, even if it improves my privacy.
56	I feel like the filtered photo reflects my gender identity.
57	I trust that the filtered photo will protect my identity online.
58	I believe that the filter will prevent an algorithm from recognizing the face in the photo.
59	I think attempts to disrupt facial recognition on social media are important.
60	I think the filter will help me avoid people who wish me harm online.
61	The filter makes me look older in the picture.
62	I feel like the filter defeats the purpose of sharing it.
63	I share photos to present myself online and this filter would not let me do this.
64	Even if this filter provided me with absolute privacy protection, I would not share the filtered photo online.
65	I will not be comfortable sharing this filtered photo because the image doesn't represent me.
66	I like the way the filter has modified my facial features.
67	I would use this filter with my photos, based on the changes it has made.
68	I wouldn't share this filtered photo because it would need me to explain its use to viewers.
69	I don't recognize the person in the image after it has been filtered.
70	I would use this filter on others' faces in a group photo but not my own.
71	The changes this filter makes are cute.
72	The changes this filter makes dehumanizes me.
73	The changes this filter makes eases the anxieties I would have around sharing it.
74	I think I would be recognizable in the filtered photo by someone I know.
75	I believe this filter protects my privacy.
76	The filter has made changes to my photo.
77	I would frequently use this filter for photos shared to social media.
78	I would use this filter on any photo I am concerned about being distributed online.

79	The promised protection against institutional actors (i.e., Facebook, Twitter) gives me comfort in sharing the filtered photo.
80	The changes to the photo are worth it for the privacy protections I am promised.
81	I could do better to increase my privacy than the filter.
82	I already do more than the filter to improve my privacy.
83	I would be willing to have this filtered photo added to a publicly available database.
84	If the filtered photo were to be saved to a public database, I would have no privacy concerns.
85	I need assurances of privacy protection given the filtered photo.
86	It is apparent to me how this filter protects my privacy.
87	I intuitively can see how this filter protects my privacy.
88	I would prefer to share the original photo over the filtered photo.
89	If I could, I would test the privacy claims of this filter.
90	I am skeptical of the privacy claims of this filter.
91	The promise of privacy protection alone gives me confidence to share this filtered photo.
92	I would be more willing to share this filtered photo if I had more control over how the image was modified.
93	I would believe the privacy claims more if the filtered photo was changed more.
94	I feel this filter addresses my privacy concerns in sharing the photo.
95	I am happy with the filtered image.
96	The changes made by the filter make me consider sharing more photos like this publicly.
97	I would use this filter when publicly posting photos I don't have privacy concerns about.
98	Using this filter would make the photos I shared with it more conspicuous than others.
99	The filtered image better reflects how I would like to be seen.
100	The filtered image looks to be more feminine than the original.
101	The filtered image looks to be more masculine than the original.
102	I don't feel comfortable with the changes made to the photograph.
103	I would want the option to select the faces I want to use the filters for in a group photograph
104	The filter is applied on the background of photo
105	I would want to blur the background of the photo in addition to faces for privacy protection while sharing it socially
106	I am confident that applying these filters would protect my privacy.

D SURVEY FORMAT



E PREPROCESSED IMAGES

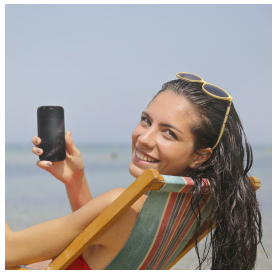
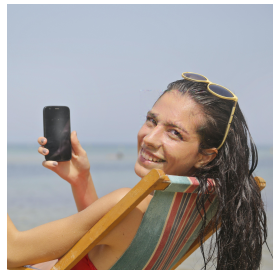
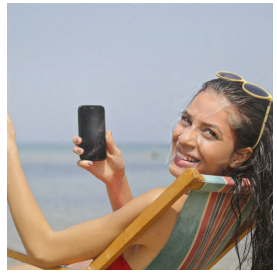
P.	Original	Pixelate	Fawkes	Lowkey
1				
2				
3				

Table 4. Preprocessed Examples

P.	Original	Pixelate	Fawkes	Lowkey
4				
5				
6				

Table 5. Preprocessed Examples (cont'd)

F PROXIES

F.1 IUIPC

F.1.1 Risk Beliefs.

(1) In general, it would be risky to give (the information) to online companies.
(2) There would be high potential for loss associated with giving (the information) to online firms.
(3) There would be too much uncertainty associated with giving (the information) to online firms.
(4) Providing online firms with (the information) would involve many unexpected problems.
(5) I would feel safe giving (the information) to online companies.

F.1.2 Collection.

(1) It usually bothers me when online companies ask me for personal information.
(2) When online companies ask me for personal information, I sometimes think twice before providing it.
(3) It bothers me to give personal information to so many online companies.
(4) I'm concerned that online companies are collecting too much personal information about me.

F.1.3 Unauthorized secondary use.

(1) Online companies should not use personal information for any purpose unless it has been authorized by the individuals who provided information.
(2) When people give personal information to an online company for some reason, the online company should never use the information for any other reason.
(3) Online companies should never sell the personal information in their computer databases to other companies.
(4) Online companies should never share personal information with other companies unless it has been authorized by the individuals who provided the information.

F.2 Uncanny Valley

Pic reference: https://stock.adobe.com/search?k=profile&asset_id=364211147

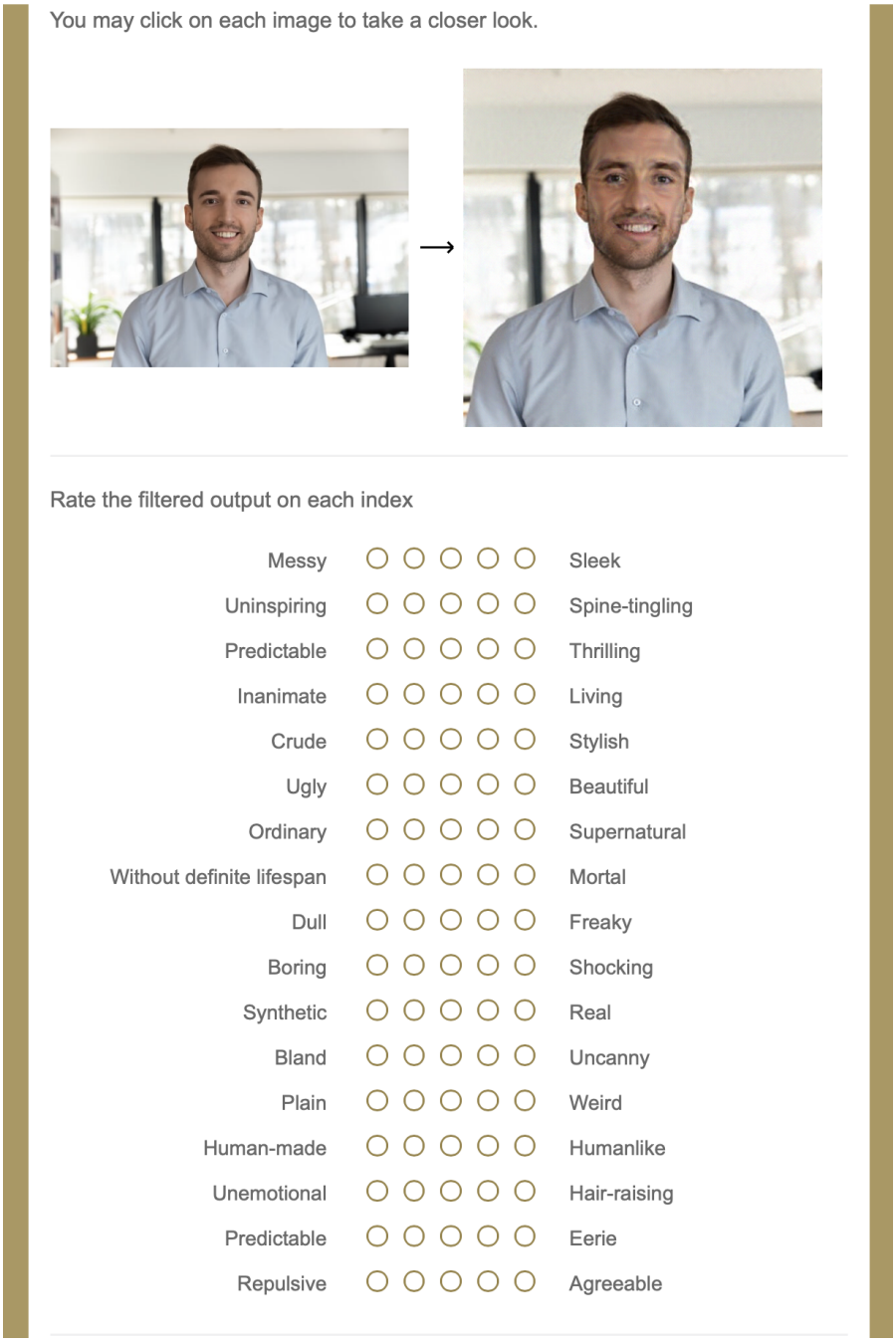


Fig. 7. Screenshot of the survey form depicting uncanny valley measure against unfiltered and filtered images[63]

F.3 Image Quality

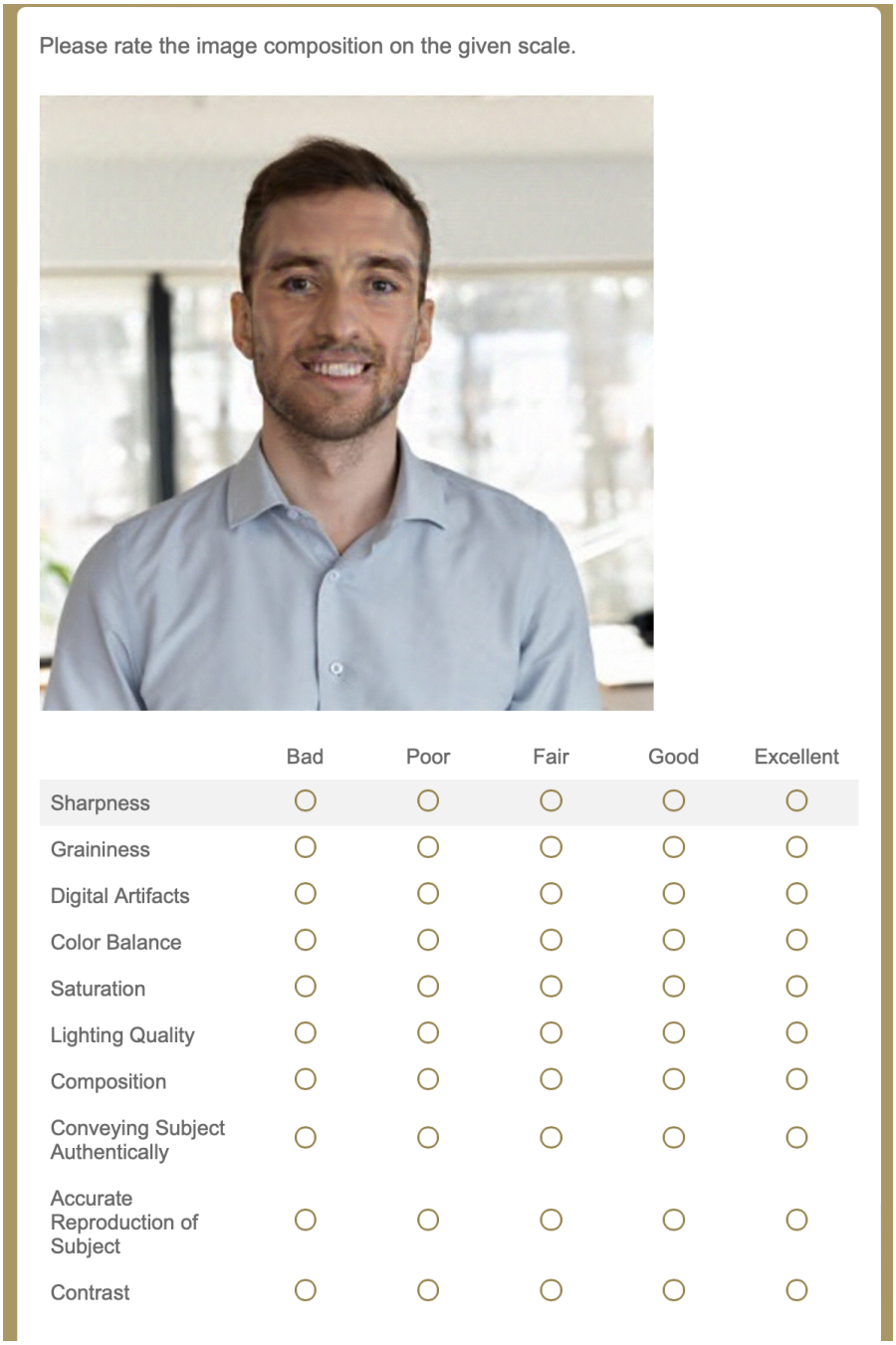


Fig. 8. Screenshot of the survey form depicting image quality measure against filtered image[76],[117]

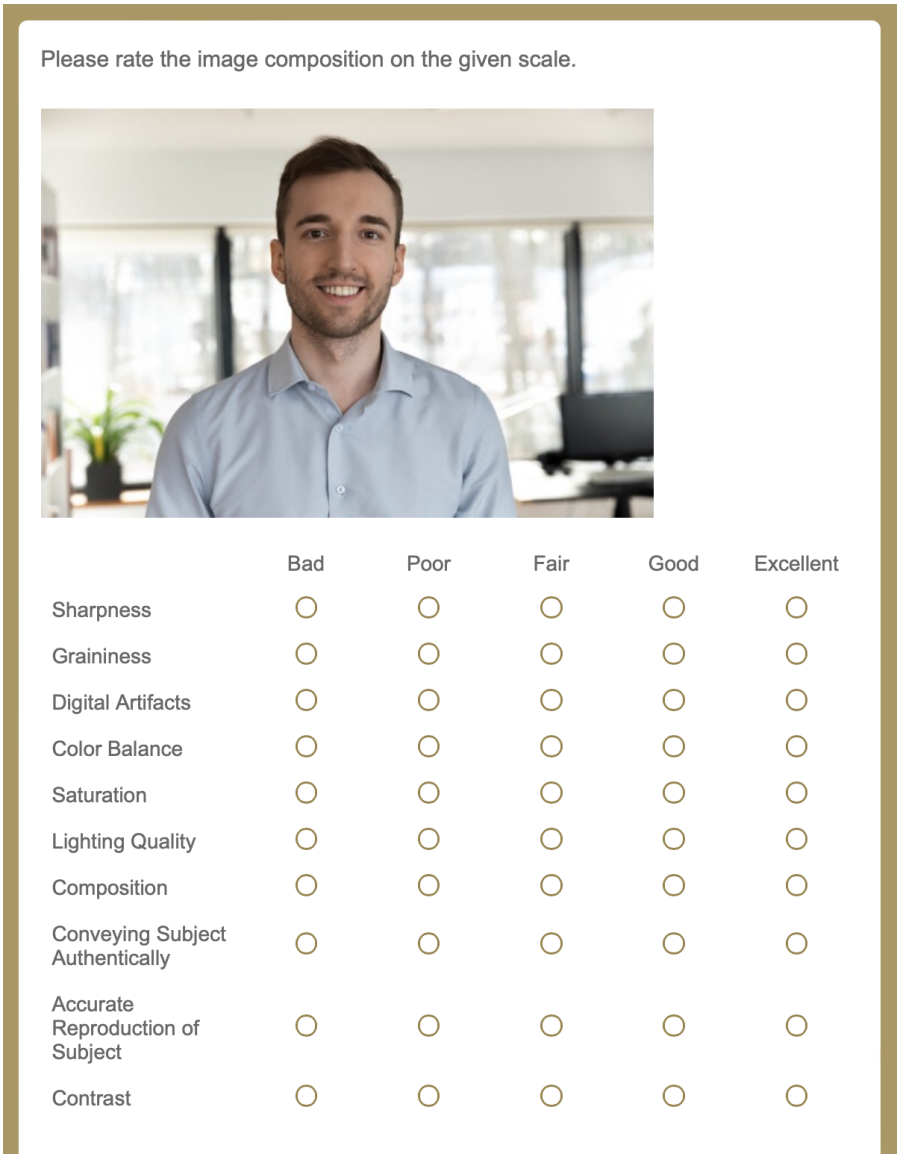


Fig. 9. Screenshot of the survey form depicting image quality measure against unfiltered image [76],[117]

F.4 Online Photo Sharing

	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
To share common interests with others	○	○	○	○	○
So that I can feel closer to others	○	○	○	○	○
To cheer myself up	○	○	○	○	○
To feel a sense of community	○	○	○	○	○
As a leisure activity, a way to relax	○	○	○	○	○
To show my friends and family what I am doing in my daily life	○	○	○	○	○
To get feedback on my photos	○	○	○	○	○
To reach a wide audience with my photos	○	○	○	○	○

Fig. 10. Screenshot of the survey form depicting online photo sharing measure against filtered image [94]

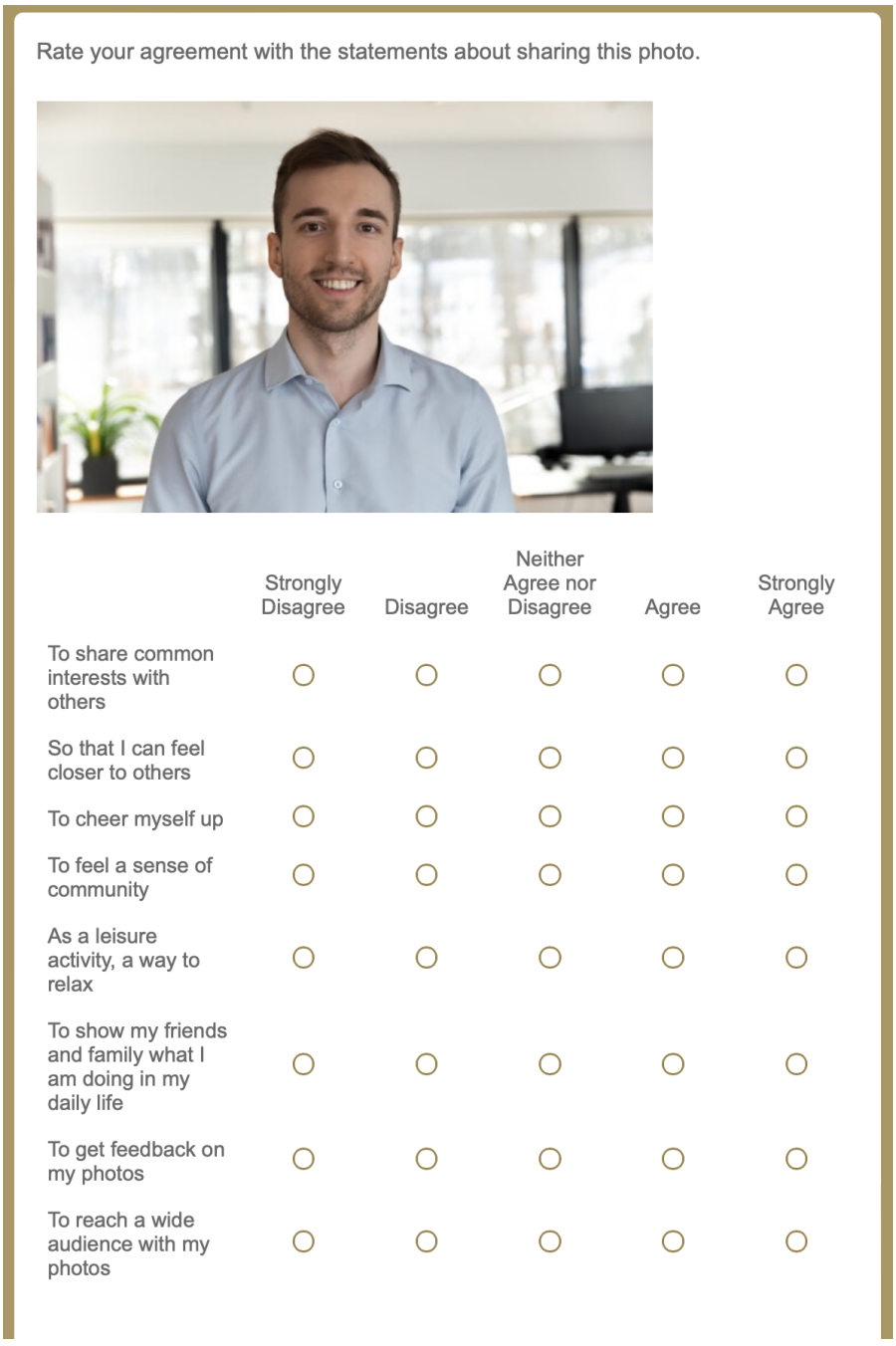


Fig. 11. Screenshot of the survey form depicting online photo sharing measure against unfiltered image [94]

Received July 2023; revised October 2023; accepted November 2023