"My face, my rules": Enabling Personalized Protection against Unacceptable Face Editing

Zhujun Xiao
zhujunshaw@gmail.com
University of Chicago

Jenna Cryan
jennacryan@uchicago.edu
University of Chicago

Yuanshun Yao
ysyao@uchicago.edu
University of Chicago

Yi Hong Gordon Cheo
gcheoyh@uchicago.edu
University of Chicago

Yuanchao Shu*
ycshu@zju.edu.cn
Zhejiang University

Stefan Saroiu
ssaroiu@microsoft.com
Microsoft

Ben Y. Zhao
ravenben@cs.uchicago.edu
University of Chicago

Haitao Zheng
htzheng@cs.uchicago.edu
University of Chicago

ABSTRACT

Today, face editing is widely used to refine/alter photos in both professional and recreational settings. Yet it is also used to modify (and repost) existing online photos for cyberbullying. Our work considers an important problem: "*How can we support the collaborative use of face editing on social platforms while protecting against unacceptable edits and reposts by others?*" This is challenging because, as our user study shows, users vary widely in their definition of what edits are (un)acceptable. Any global filter policy deployed by social platforms is unlikely to address the needs of all users, but hinders social interactions enabled by photo editing.

Instead, we argue that face edit protection should be implemented by social platforms based on individual user preferences. When posting an original photo online, a user can choose to specify the types of face edits (dis)allowed on the photo. Social platforms use these per-photo policies to moderate future photo uploads, i.e., edited photos containing modifications that violate the original photo's policy are either blocked or shelved for user approval. Realizing this personalized protection, however, faces two immediate challenges: (1) how to accurately recognize specific modifications, if any, contained in a photo; and (2) how to associate an edited photo with its original photo (and thus the edit policy). We show that these challenges can be addressed by combining highly efficient hashing based image search and scalable semantic image comparison, and build a prototype protector (*Aletheia*) covering nine edit types. Evaluations using IRB-approved user studies and data-driven experiments (on 839K face photos) show that Aletheia accurately recognizes edited photos that violate user policies and induces a feeling of protection to study participants. This demonstrates the initial feasibility of personalized face edit protection.

*Work completed while at Microsoft

Proceedings on Privacy Enhancing Technologies 2023(3), 1–16

https://doi.org/10.56553/popets-2023-0066

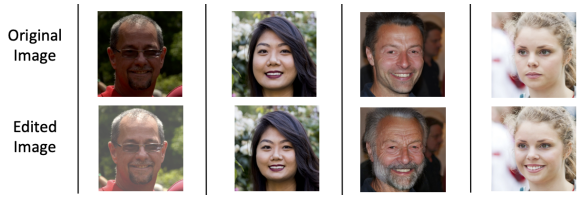

Figure 1: ***Examples of face edits done by today's low-cost or free edit tools (Photoshop, PortraitPro, FaceApp).***

KEYWORDS

face edit, personalized protection, image moderation

1 INTRODUCTION

How we are perceived online can be heavily influenced by images of our faces online. To achieve a desired presentation, many users prefer to have their face images digitally edited or refined before posting online [19]. Popular photo sharing sites, social networks, and mobile apps now allow users to easily edit faces in images for a variety of uses, including beautification, collaboration, and recreation. With a single button, users can touch up a face photo, change the person's age, face shape, expression, and other facial features. These edits are so realistic that it is difficult to identify originals from edited versions with the naked eye. Figure 1 shows four photos along with their realistic edits.

As face editing tools grow more common, however, negative impact from misuse and abuse also grows. For example, one widely known threat is "Snapchat Dysmorphia" [53], where many edited selfies reach unrealistic beauty standards, changing how young people look at themselves, leading to low self-esteem and mental health issues. Our work looks at a second, equally harmful threat: "misuse of face editing on someone's online photos as an effective form of cyberbullying" [31, 41, 59]. This takes place when others download online photos of a user, edit them, and repost them for malicious purposes. For working adults, photos from LinkedIn might be edited and reposted to discredit or harass someone in the workplace. For younger users, photo editing is already used for cyberbullying [41], a problem experienced firsthand by a majority of teens [6, 20, 24]. While platforms are aware of photo editing as a

tactic in their efforts to curb bullying [44, 59], experts predict photo editing tools are likely to become "a potent tool of cyber-stalking and bullying" [29]. These give rise to the following question:

"*Can we support the collaborative use of photo editing and sharing, while protecting against unacceptable editing and resharing of photos that we have posted online*?"

When studying this problem, we identify two key issues. First, there is a lack of understanding of how users perceive photos edited by others. Existing works studied how users edit their own photos (e.g., [60]) but not their attitudes towards edits and reposts of their photos done by others. Second, there is a lack of protection tools against harmful photo edits and reposts. Existing works focus on controlling viewership [45, 56, 58] or obfuscating sensitive content like faces [23, 35], but do not protect users against new uploads containing unacceptable edits of their online photos.

With these in mind, our work begins with a user study that explores how users perceive others editing their face photos. We find that users vary significantly in their tolerance, depending heavily on the type of edits. The results indicate a clear need for "personalized photo moderation tools" that protect users against wrongly edited images. This task is challenging, because such moderation tools must walk a fine line between reliably detecting unacceptable face edits based on individual user policies, and overly sensitive filters that hinder social interactions enabled by photo editing/sharing.

Today's content moderation tools fall far short of these goals. To date, research on image analysis largely focuses on the problem of detecting *whether* a face photo has been manipulated (e.g., [13]) rather than how it has been edited. This is motivated by detection of deepfakes [64], often in the context of AI-generated fake photos/videos that misrepresent public figures or fabricate news events. These detectors assume that there are no available "originals" of these photos, and mainly operate by detecting specific digital artifacts left on the photo by deepfake tools. In our problem context, these detectors often fail to even detect the fact that a face has been edited, much less the type of face edit or whether it is acceptable under a specific user policy (see Table 4 in §7).

These findings motivate us to propose, design and prototype *Aletheia*, a new photo moderation tool that protects individual users against the uploading/sharing of unacceptably edited versions of their online face photos. Photo hosting services and social media networks can deploy Aletheia to protect their users and their posted face photos, where owners of original photos can now specify their willingness to allow or disallow any categories of edits to their online face photos. For example, a user **U** on Microsoft OneDrive may allow facial retouching on their personal photos, but not changes in age. A bully **B** trying to share an age-changed photo of **U** is detected by Aletheia as violating **U**'s face edit policy, and the upload is flagged for moderation or rejected depending on **U**'s policy. Such personalized protection provides each user with full agency in deciding how their online photos can be edited, which we hope will stimulate healthy social interactions enabled by face edit tools while protecting users from their misuse.

Our Contributions. Our work targets the critical challenge of user-specified moderation of how others make face edits on our online photos and repost them. Our work makes three contributions:

(1) a user study to explore user tolerance of face edits on their photos when done by others. Our results show significant variance across users and edit types, motivating the need for *personalized* face edit protection;

(2) Aletheia, a prototype moderation tool for photo sharing sites to implement user-specific face edit protection on their photo posts. We address the key challenge of recognizing the type of edits contained in a photo x by combining highly efficient hash based image search that locates x's original, unedited version, and scalable semantic image comparison between x and its unedited version. This "reference-based" methodology differs from existing solutions for detecting deepfakes, which assume the absence of an original image. We plan to release Aletheia for academic use;

(3) IRB approved methods (user studies& data experiments on 839K photos) to evaluate Aletheia for protection effectiveness, scalability, and users' perceptions of Aletheia's protection on their online photos. Results show i) Aletheia successfully identified 93.8% of edited photos marked as unacceptable by user study participants; ii) Aletheia operates at scale with high accuracy and low latency, e.g., > 97% accuracy and <1s latency per photo in detecting edited images, while existing works offer only 9.5 − 55.6%; and iii) study participants had generally positive views of protection provided by Aletheia. Altogether these results suggest that our approach could be an effective method for protecting online face photos from being improperly edited and reposted.

Finally, we also discuss current limitations and future directions to push the concept forward.

1.1 Broader Impacts and Ethics

Our proposed design allows online platforms to support social interactions via photo editing and sharing, while giving users agency over how their photos posted online can be altered by others. Our goal is to bring attention to this important problem, and to spur efforts by providers to discourage and reduce potential abuse of (face) editing tools.

Potential for Misuse. Aletheia could potentially be misused in two ways. In one scenario, someone could perform a "denial-of-service" attack on a user **u** by uploading many improperly edited copies of **u**'s photos, triggering the system to send **u** many alerts and thus injecting stress and mental load of reviewing edited images on **u**. To resist this type of misuse, **u** can opt for more automated policies, or relying on stronger upload rate limits by the provider.

In another misuse case, someone can block **u** from uploading their own images by uploading an edited version first, claiming it as the original, and applying a strict, no-edit policy. These conflicts can be identified by allowing a denied upload request to be challenged, where **u** can claim ownership of the photo by verification via face recognition or manual inspection, or leveraging hardware-generated stamps [38]. On the other hand, this aligns with the well-known issue of ownership and copyright of online photos, which has been a topic of much debate and discussion [43]. Ideally, only the legal owner of an original photo should be the one who defines the edit policy in a system like Aletheia.

Overhead. Processing overhead for Aletheia will scale with the volume of uploads for larger photo-sharing platforms, particularly if integrated as a collaborative system across multiple platforms.

We expect performance to significantly improve in followups to this initial proof of concept.

Ethics. We took careful steps to protect user privacy throughout our study. Our evaluations were vetted and approved by our local IRB council. All original face images used in our user studies and shown in this paper come from the Nvidia FFHQ dataset [46] and the Google DeepFake Detection dataset [16, 49], under licenses explicitly granting free use for non-commercial research and educational purposes. Other face images used in our experiments were obtained from public datasets under agreements that grant non-commercial academic use. Images were stored in a secure server at our institute and only used by the authors to evaluate the accuracy of Aletheia's face edit detection/recognition.

2 BACKGROUND

We now provide background for our work on face edit protection, and discuss related work on online photo privacy and protection.

2.1 Misuse of Face Editing on Others' Photos

An alarming and growing trend is editing and reposting of other people's selfies posted online, without permission and often with intent to harass and bully individuals [48]. Incidents such as students posting unappealing edited photos of others [41] have contributed to photo editing being listed as a popular cyberbullying tactic [59].

Face-edits vs. Deepfakes. Deepfakes involve high quality *fictional* images or videos heavily edited or created using algorithms, often involving deep learning models such as generative adversarial networks. On numerous occasions, bad actors have leveraged deepfakes to generate and disseminate malicious videos of public figures (e.g., House Speaker Nancy Pelosi) or sway public opinion with fake news [5]. Many ongoing efforts attempt to detect deepfakes [4], measure their effects and dissemination [26].

While deepfakes typically focus on depicting fictional actions or events, face-edits are manipulations of existing images. Deepfake research broadly focuses on identifying *if* deepfakes are real (or synthetic), while our work focuses on identifying *how* face images have been edited, and whether they are acceptable edits based on personalized policies.

2.2 Online Photo Privacy

User Perception of Online Photo Privacy. While most users are concerned with online privacy, they vary greatly across users, along with the actions (if any) users decide to take [12, 14]. A user's view of privacy is affected by their personal traits (e.g., age, gender appearance) and their own definitions of privacy [47]. Users also vary significantly in their definitions of harassment, complicating enforcement of anti-harassment policies online [9, 10].

Protecting User's Photo Privacy Online. Existing strategies can be broadly categorized into two approaches [36]: 1) deploying policies to control viewership or moderate content, and 2) adding artifacts onto images to obfuscate privacy-sensitive content.

Within the first approach, existing efforts have proposed multiple methods of policy management. These include strategies to configure per-user privacy settings [45, 56] or manage collaborative privacy settings among users [8, 58], visualization tools to explain privacy settings to users [17], and systems that employ human users to moderate content [18].

For the second approach, existing works have developed obfuscation techniques (e.g., blurring, distorting or blocking) to protect sensitive content on an image (e.g., scene element, face) [35]. While obfuscating a face before uploading a photo could effectively prevent others from editing the face, doing so adversely affects viewers' experiences [23], defeating the original purpose of sharing photos online [36]. Some recent works propose to add artifacts/perturbations to images, so that they interfere with certain photo edits produced using deep learning models (i.e., faceswap [52]). However, these are highly customized towards specific types of edits, and must be applied to images before sharing.

2.3 Image Moderation Tools

There is considerable effort by security and computer vision research communities to develop techniques that detect the presence of edits in photos. They can be broadly divided into three categories.

Embedding and verifying image signature. This solution seeks to provide integrity guarantees of an image via cryptographically-secure digital signatures. Such signatures would be applied to images at their creation (e.g., by smart cameras), and validated by the consumer (e.g., digital photo frame). However, this approach severely restricts the flexibility needed by content editors who need to refine and edit images before it is ready for consumption.

Examining image edit logs. Photo edit tools can log specific edits made to each image into the image's metadata. For example, a recent Adobe proposal calls for its tools to embed edit logs into images as a secure hash [3]. An online service can extract the edit log from images and reject uploads of images whose logs contain unauthorized edits.

Inspecting image visual content. Numerous graphics-based and deep learning-based tools try to detect face edits in a target image by studying its visual content. These are generally referred to as image manipulation detection and/or deepfake detection. These approaches are *reference-free*, i.e., they operate *directly* on the target image and do not assume knowledge or access to the original image. Along this line, existing works mostly target specific types of edits (e.g., faceswap [33, 64], image splicing [25]) or a specific tool (e.g., adobe photoshop [62]). They function by detecting the presence of edits based on digital artifacts introduced during face editing. These include visual artifacts (e.g., resolution inconsistency [34], temporal inconsistency [21]), behavioral anomalies (e.g., inconsistent head pose and expression [4], abnormal eye blink pattern [33]), and DNN model artifacts (e.g., GAN fingerprints [65]).

3 UNDERSTANDING HOW USERS PERCEIVE FACE EDITS DONE BY OTHERS

Our discussion in §2 shows that despite existing studies on self photo editing, deepfakes, and photo privacy, there is little work on understanding users' perspectives and reactions on face edits that others have applied to their online photos. To answer this question, we conducted an online survey about users' tolerance for others editing their selfies and perceptions of privacy when posting photos online. Our study was approved by the local Institutional Review

Board (UChicago IRB-20-1230). The full survey script is available at http://sandlab.cs.uchicago.edu/faceedit/userstudy.

3.1 Our User Study Design

Participants. We recruited 100 participants via the crowdsource platform Prolific (https://www.prolific.co/). Participants were required to be 18+ years old, live in the US, and have 95% approval rating on Prolific. The survey was designed to take 15 minutes on average, and participants received $3 as compensation. We collected 99 valid responses (one participant timed out), among which 53 identify as male (46 female). The age distribution is 18-29 years (60%), 30-39 (22%), 40-49 (16%) and 50-59 (2%).

Task. We first presented the concept of face editing to participants, and asked whether they have observed face edits done by others in online images/videos (not necessarily their own). We then asked participants to suppose they shared a photo of their faces online to a platform similar to Facebook or Instagram, and asked a series of multiple choice and free response questions about their perceptions and opinions regarding others editing the posted photo and reposting it (e.g., what edits can/cannot be tolerated), and their preferences for how the platform should act regarding these edited versions. For our study, we categorized common types of face edits (offered by today's tools) into five groups by their effects [13], from which we produced 15 edit types (see Table 12) used for our study.

The goal of our user study is not to develop (or apply) a method to *precisely* collect a per-photo edit policy from participants, but to explore the pattern and diversity, if any, in participants' responses to others applying face edits to their photos.

Conditions. We presented two scenarios (in random order): the edited photo is viewable to *friends and family only* (similar to Instagram's "close friends" option) or viewable to the *public*. For each scenario, we surveyed participants in two steps. In step 1, we described each edit type and then illustrated its high-level effects using example photos (we explain the photo choices below). We then asked each participant to imagine such edit type (with varying spectrum and style) is applied to their own photos by another person, and rate how likely they would allow the edited photo. The default rating is 5-point Likert scale, i.e., never (1), rarely (2), sometimes (3), usually (4), always allow (5). For edits (e.g., age, brightness, face shape) that can be measured on a spectrum (e.g., increase/decrease), we also presented examples of five edit levels (0%, 50%, 100%, 150%, no limit) on both increase and decrease, and asked participants to which extent they would allow the edit. Next in step 2, we presented several new edited images and asked participants to select the ones they would allow. To verify responses, we included attention check questions and applied both time check and manual inspection to detect straight-lining and false input. Appendix B lists examples of survey images.

Photo Choices. To precisely collect a participant's opinion on face edits, one could present them samples of edited photos of themselves. However, seeing certain edits on their own photos could lead to negative emotional effects that cannot be predicted before the study [31, 40]. Also the remote/one-direction nature of our user study meant we could not debrief our remote participants. Thus to minimize potential harm, we did not collect or alter personal photos from our remote participants. Instead, we showed participants sample photos of other people, before-and-after edits, to help illustrate possible effects of different edit types. We asked participants to visualize edits applied to photos of their own faces when answering the study questions. In our opinion, doing so achieves the desired goal of impressing the impact of different face edit types to individual participants while minimizing any potential negative emotional effects on them.

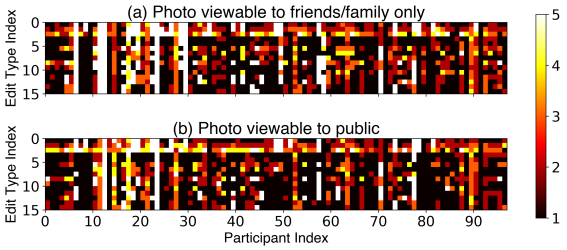

Figure 2: ***The raw score distribution across our study participants (99 users), who provided a score (1-5) for each of the 15 edit types. 5 = always allow, 1 = never allow.***

3.2 Key Findings

Finding #1: Face edits by others are commonly observed in today's online platforms. Most participants (75%) reported having observed face edits done by others in online shared images/videos. When asked about how frequent they observe such editing, 31.3% reported 'Somewhat often (a few times a week)' and 12.1% reported 'Very often (at least once a day)'. Also, 19.2% indicated that they themselves have edited other people's face images/videos.

Finding #2: Users vary significantly in tolerance for different types of face edits. Participants showed significant variation in their tolerance of others editing their online face photos. This can be observed from the raw scores on the edit tolerance (rated on a scale of 1-5) in Figure 2, where we show the scores of all 99 participants for each of the 15 edit types. Here the color represents the raw score, white = 5 (always allow) and black =1 (never allow). On average, participants would allow half of the editing types presented, with 8 participants (8%) allowing *all* types of edits (i.e., a score of 5 for all 15 types) and 3 participants (3%) allowing *none* for either scenario (i.e., a flat score of 1). We also measured the level of variation as the standard deviation (std) across edit types and participants. For each edit type, the std across participants is high and comparable to the mean (std $\in [1.17, 1.58]$, mean $\in [1.62, 3.4]$). For participants allowing some edit types, the std across edit types is similarly high (std $\in [0.25, 2.0]$, mean $\in [1.07, 4.0]$), suggesting that their choices of the edits are highly personalized.

To explore the impact of context (i.e., photo viewable to public or friends/family only), we computed the difference between the mean tolerance of two scenarios per participant. Again the results vary across participants: 52.6% showed indifference, 20.6% would allow more edits for public view, and 26.8% would allow more edits for friends/family view.

To understand the reasoning behind their individual selections, after evaluating both scenarios, we asked participants to explain in their own words. As shown in Table 1, the reasons expressed fall

Reason	Example
None/ Very Few Edits	"I **wouldn't want anybody editing my photos, whether I know them or bot** [*sic*]. It feels intrusive."
Specific Edits Only	"It doesn't matter to me who can see it, I **just don't want specific edits** done to me." "Sometimes **some edits end up making the pictures weird**"
Allow More Edits among Friends/ Family	"Well I would find it **funny if my friends did some of those edits** to me **but I would be a bit annoyed if a random person** did some of those to my photo."
Allow More Edits for Public View	"would have **more fun doing more extreme edits for public viewing** because potentially more people will be seeing them rather than friends and family who already know what I look like."
General Indifference	"I feel like **whatever you show in private for the most part you should be able to show in public**"

Table 1: *Participant reasons for their edit preferences.*

into 5 general categories: prefer no edits ever, prefer only specific edits (regardless of the audience), would allow more edits among friends/family, would allow more edits for public photos, or general indifference. These responses also indicated that who the editors are is also an important factor. Overall, we can clearly observe that users differ largely in their tolerance of face edits.

Finding #3: Many users prefer aggressive identification and notification of policy violations. Our study showed that participants preferred a more aggressive approach to detecting unacceptable face edits. Two-thirds of participants felt the platform should **flag as many potentially edited images** as possible, even at the cost of some false positives. When the system detects an edited image that violates user preferences, 87% of participants wanted proactive notifications. Finally, 60% of participants expressed concern about the development of new face-editing methods, and the need to adjust their preferences accordingly over time.

The need for personalized face edit protection. Overall, our user study shows that users are heavily concerned about others editing and reposting their face photos and want the ability to protect their online photos; but since users hold very different definitions of what face edits are unacceptable, the protection against face edits must be *personalized*.

4 PERSONALIZED FACE EDIT PROTECTION VIA IMAGE MODERATION

We believe a viable solution for personalized face edit protection would involve online photo platforms that deploy "photo moderation tools" to monitor photo uploads. When a platform detects a new photo upload containing edits that violate the preferences of the original photo's owner, the platform will block or tag the photo based on user preferences. However, developing such a moderation tool is challenging. Given an image to be inspected, the tool must not only detect *if* the image is an edited photo, but also *how* it was edited. As we explain below, current image content moderation systems/tools fall far short of these goals.

Existing moderation methods are insufficient. As summarized in §2.3, there exists considerable efforts by security and computer vision researchers to develop moderation techniques that detect face edits in photos. Here we discuss why they are insufficient for the task of personalized face edit protection.

(1) *Verifying embedded signature*: This approach embeds a digital, cryptographically-secure signature into an image, so that any edit that destroys the signature can be detected [30]. However, existing tools provide a binary answer (i.e., edited or not), and cannot be parameterized to detect specific types of edits.

(2) *Examining image edit logs*: Some edit tools can log specific edits into the image's metadata (i.e., Adobe Content Authenticity Initiative). Online services can extract the edit log and reject images with unauthorized edits. However, this only works if *all* edit tools consistently and reliably preserve these metadata fields, which is an unrealistic assumption.

(3) *Inspecting image visual content*: Many propose to detect face edits by studying image visual content and identifying digital artifacts introduced during face editing [4, 34]. However, these methods remain unsuitable for our task because: (1) they focus on *detecting* the presence of face edits rather than *recognizing* them; (2) they rely on artifacts of specific face editing, and thus do not generalize across edit types and tools; (3) continued efficacy of such detectors is in question, because face edit tools are evolving to reduce or completely eliminate digital artifacts. Later in §7, we evaluated three state-of-the-art detectors, and find they often fail to detect a face has been edited at all.

5 ALETHEIA

To address the unfulfilled need for personalized face edit protection, we propose and design *Aletheia* to address this gap. Here, we present the goals, assumptions and design intuition behind Aletheia. In §6 we present a detailed design of its two core technical components.

5.1 Goals and Assumptions

Aletheia is an image moderator system to protect original face images on photo sharing services[1]. It is designed to:

- allow users to specify (and update) their personalized policy on unacceptable face edits on their original photos;
- identify images containing unacceptable edits and trigger subsequent actions defined by the policy.

Usage Scenarios. Here we make two assumptions:

- Aletheia focuses on selfie photos (front-shot of a single face), which are the main target of malicious face editing.
- We design Aletheia to protect a user's face photos after they are posted online. To receive protection, an original photo must be registered into Aletheia before its edited versions. Specifically, when a user posts an original photo into an online service employing Aletheia, the photo is verified by Aletheia as an original and then registered into the system. A user can fill a claim with Aletheia if their original images are registered by someone else, and prove ownership by verification via face recognition or camera-generated stamps [38].

Threat Model. We are motivated by the need to prevent the use of face edit tools for cyberbullying, and design Aletheia to resist "standard manipulators" who are familiar with everyday technology (i.e., those who can use commodity tools to modify photos, and delete/modify a photo's metadata), but not security

[1]Multiple services can cooperate to expand the protection coverage. In this paper, for simplicity, we consider a single service.

experts or strongly motivated adversaries (i.e., resourceful attackers who analyze Aletheia's internal design and craft adversarial attacks to bypass Aletheia's detection). Later in §8, we perform a security analysis on Aletheia against those strong attackers who carefully craft face edits to evade detection, and outline potential defenses.

5.2 Design Intuition

Different from existing efforts, we design Aletheia to effectively detect if and how an image has been edited, by applying *reference-based* face edit detection and recognition.

Reference-based face edit recognition. Aletheia recognizes potential face edits on x by comparing pairs of images: x and its original/unedited version x_0 that Aletheia will locate. Upon receiving a request to upload a face image x, Aletheia first identifies whether x is an original or edited photo. If x is edited, Aletheia locates its original version x_0 from its database of original faces, and compares x_0 and x to identify any unacceptable face edits specified by x_0's policy.

Key benefits. Aletheia presents three key advantages over existing content moderators:

- allowing users to define personalized protection rules for their original face images x_0.
- transforming the extremely challenging problem of recognizing types of edits in a single face image to a manageable problem of recognizing differences of two images.
- remaining agnostic to edit tools and scaling to new edit types, because Aletheia recognizes face edits by extracting and comparing semantic face attributes (e.g., age, expression) of the original and edited images.

5.3 System Architecture and Dataflow

Aletheia consists of four components: **(1) a face edit policy manager** that allows each user, when uploading an original face photo, to specify their policy that defines unacceptable face edits and the subsequent system action upon detecting such edits; **(2) an image inspector** that for each incoming image x, inspects the image to determine whether it is an original image; if so, the inspector asks the user to input policy, and if not, it locates x's original version x_0; and **(3) an edit recognizer** that compares x and x_0 to determine whether x contains any unacceptable edits defined by x_0. In addition, Aletheia maintains an internal **(4) database** to store registered original face photos and their edit policy.

Figure 3 illustrates Aletheia's operation pipeline for two scenarios. In scenario I, the input image to Aletheia is an original face photo. The image inspector verifies the input is original, prompting the user to define an edit policy on this photo via the policy manager. It then registers the photo (and policy) into the database, and accepts the upload request. In scenario II, the input image is an age-edited face photo. The image inspector first identifies the input as an edited photo and proceeds to locate the original face photo (and edit policy) in the database. Then the edit recognizer compares both photos to identify edits, and uses the edit policy to determine existence of any unacceptable edits. If so, the upload request is either rejected or flagged for user review (per user's policy). If not,

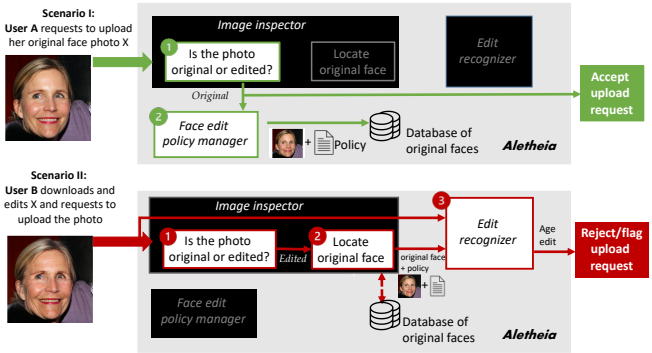

Figure 3: ***Overview of Aletheia's operation when users request to upload original (scenario I) and edited photos that contain unacceptable edits (scenario II).***

the upload request is accepted. In the example of Figure 3, the image violates the user policy that disallows age edit.

Edit policy specification. By recognizing differences between x and x_0, Aletheia can support a flexible configuration of edit policy per user/image. For an original image x_0, the owner can specify one or more edits (from a list provided by Aletheia, or self-defined) to disallow if the amplitude and/or direction of the edit exceed some thresholds. For example, one can disallow an age edit if the edited face appears as more than 50 years old, any edit that changes the expression, or any edit that changes the skin tone index.

Edit recognizer customized by user edit policy. After an input image x is flagged as edited, Aletheia's edit recognizer examines x based on the edit policy of its original copy x_0. Given the set of edits marked as unacceptable by x_0, Aletheia verifies whether any is present on x and exceeds the threshold defined by the policy. Once a violation is detected, the owner of x_0 can choose to reject x or review x themselves (per Finding #3 in §3). We believe this provides users agency over how others can alter their photos.

Design focus. Focusing on exploring the feasibility of Aletheia, we consider in this work a simple policy design – users choose types of edits to disallow from a list provided by Aletheia and define a threshold based decision rule per edit. Clearly, Aletheia would benefit largely from an interactive interface to provide clear interpretation of face edits, guide users in defining their policy, and translate the policy into rules that Aletheia can implement. We leave this to future work (see §9).

6 DETAILED DESIGN OF ALETHEIA

We present the detailed design of image inspector and edit recognizer. The two hold different goals: image inspector decides whether an image is original or edited, rather than how to recognize edits.

6.1 Image Inspector

For an incoming image x, Aletheia determines whether x is an original face photo or an edited copy; if x is edited, locates its original version x_0. For this, we propose a 2-step process to boost accuracy while lowering computation cost.

Step 1: Estimating a photo's edit status using "image provenance". Aletheia first applies a "provenance" based method to

quickly "estimate" the edit status of x, i.e., original or edited. This leverages the fact that the image hosting service running Aletheia, when publishing any image on the service, can embed some provenance data (i.e., a string identifying the image's original copy if it is an edited copy) into the image's metadata[2]. As a result, any original image published by the service will contain an empty provenance field, and any edited image published by the service will contain a provenance field identifying its original copy. Assuming that normal use or edit do not remove the provenance data, Aletheia can simply inspect the provenance data in an image to "estimate" its edit status. We note that removing or modifying these metadata by each user is possible but requires manual efforts or specific tools. None of the 10 common editing tools considered by our work remove or modify the metadata field.

Addressing empty or manipulated metadata. On the other hand, some online services, such as Facebook, Twitter, and Instagram, choose to delete the metadata of an uploaded image in order to protect user privacy. As a result, online photos posted on these services will have an empty provenance. Similarly, a user can also intentionally or accidentally remove or modify an image's metadata before posting it, so that the image either "claims" to be an original, or "claims" to be the edited version of another photo (e.g., the one allowing any edit). Aletheia addresses both scenarios by applying a verification step after Step 1 to verify the "original" status of an image declared by its metadata (Step 2a), or the original copy of an edited image (Step 2b).

Step 2a: Verifying "original" photos by hashing based image search. Upon receiving an image with empty provenance data, Aletheia runs a verification program that compares the image's visual content with those stored in the original face database. Intuitively, a new original face photo should be reasonably different from those stored in Aletheia's database of original photos. That is, the image's minimum perceptual distance to those in the database should be higher than some threshold. Such distance can be computed by a database-level image search/comparison.

We designed our inspector to realize this concept, and more importantly, to address two key challenges in practical deployment. First, the mass scale of today's photo platforms makes the database-level image search intractable if we compare images in the raw pixel level. Instead, Aletheia applies perceptual hashing to convert each image's content into a single compact hash (e.g., 64 bit) where the hamming distance between hashes well approximates the perceptual distance between images. These perceptual hashes are compact and fast-to-compute, making them a good fit when searching over hundreds of millions, or even billions, of images [11]. Second, to flag edited photos that contain large changes, Aletheia builds two content representations (whole-image, background-only) to expose similarity between an original photo and its edited versions. For each representation, Aletheia runs the hash-based, database image search to identify the candidate image most similar to the target image. This produces two candidates. Aletheia then computes the raw visual similarity between each candidate and the target image, measured by the Structural Similarity Index (SSIM) [68]. If any of the two SSIM scores is higher than a threshold θ_{SSIM}, the target image is an edited photo; otherwise, it is an original photo.

Step 2b: Verifying the original copy of an edited photo. After detecting that x is an edited face photo, Aletheia moves to locate its original copy x_0 from the database of original faces. If x has no provenance data, the previous step (Step 2a) should have already found its original copy. But if x comes with the provenance data that announces its original copy x_0, Aletheia needs to verify whether the declared x_0 is indeed the original copy. Again this verification is done in two steps. First, Aletheia checks whether x and its declared x_0 are sufficiently similar in visual content, again by computing their SSIM. If their SSIM $> \theta_{SSIM}$, the verification passes. If not, Aletheia applies the same database search in Step 2a to locate the true original copy, using the same threshold θ_{SSIM}.

Choosing θ_{SSIM}. Aletheia configures θ_{SSIM} based on the detection false positive rate. Intuitively, the choice of θ_{SSIM} should ensure that, with a very high probability (e.g., p=99%), any two distinct original photos should not be verified as a pair of an original photo and its edited version. Thus, Aletheia sets θ_{SSIM} as the top p-percentile value on the SSIM score of image pairs sampled from the database.

Speeding up image search. To speed up the image search in Step 2a, we apply a ball-tree based structure [37] to index the hashes in our database, and runs a k-nearest neighbor search for x's hash on the ball tree. This reduces the search cost to $O(dlogN)$ [37], where d is the hash dimension (i.e., 64 bit), and N is the number of images in our database of original faces.

6.2 Edit Recognizer

Given x and its original copy x_0, Aletheia applies semantic image comparison to recognize edits in x. As shown in Figure 4, Aletheia extracts from x and x_0 a set of relevant semantic attributes (e.g., those related to the unacceptable edits defined by x_0's edit policy), and compares the attributes of two. Example attributes include age, identity, facial expression, face shape, skin tone and hair color. Thus the editor recognizer has two components: (i) face attribute extractors, and (ii) user-specified decision rules per edit type.

Leveraging existing face attribute extractors. Aletheia leverages existing (and ongoing) efforts on predicting semantic attributes from face images, such as public, pre-trained classifiers on age [51], expression [7], identity [54], and face segmentation models [69]. This modular design means that Aletheia can easily replace each deployed attributor detector by a newer, more advanced version when it is available. And the performance of Aletheia depends on precision of these classifiers.

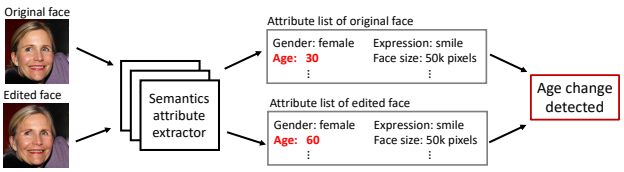

Figure 4: ***Aletheia detects the edit types by comparing semantic attributes of target image and its original copy.***

[2]The provenance data should remain intact through certain photo usage/edit. Embedding it into the metadata that already exists internally in image files is a viable solution, since this metadata is widely supported by image formats such as JPEG, DNG, PNG and TIFF, etc, and does not modify the visual content of the image.

Edit Type	**Face Attribute Extractor**	**Default Decision Rules** (x_0:original photo, x:edited photo)
Faceswap	Face Verification Model [54]	**Identity**(x_0) ≠ **Identity**(x)
Change expression	**Expression** Classifier [7]	**Expression**(x_0) ≠ **Expression**(x)
Change gender appe.	**Gender Appearance** Classifier [15]	**GenderAppearance**(x_0) ≠ **GenderAppearance**(x)
Change skin tone	Face Segmentation [69] + Skin Color Identifier	**FaceColor**(x_0) ≠ **FaceColor**(x)
Change hair color	Face Segmentation [69] + Hair Color Identifier	**HairColor**(x_0) ≠ **HairColor**(x)
Add/remove eyeglasses	Face Segmentation [69]	**EyeglassDetected**(x_0) ≠ **EyeglassDetected**(x)
Change age	**Age** Classifier [51]	| **Age**(x_0) - **Age**(x) | ≥ 5 years
Change faceshape	Face Segmentation [69]	| **FaceShape** (x_0) - **FaceShape** (x) | / **FaceSize** (x_0) ≥ 5%
Change brightness	**U**: average brightness	|**U** (x_0) - **U** (x)| ≥ 10

Table 2: ***Our prototype of Aletheia is configured to recognize 9 edit types using public models and default decision rules.***
◇ *We follow the human skintone color palettes defined by [2] to define skin color.* † *We follow the hair color palettes defined by [1] to identify hair color.*

Supporting user-specified decision rules. After extracting semantic attributes from x and x_0, Aletheia compares each pair of attributes to determine whether the corresponding edit is present on the edited photo. Here the decision metrics (and thresholds) are configurable by individual users as a part of their edit policy, leading to personalized face edit protection. For example, a user can treat 5+ years as an indicator of age change, and 5+% changes in the detected face as an indicator of the faceshape edit, while another user can define 10+ years as age change and 10+% change of face size as faceshape edit.

Detailed implementations. We list in Table 2 the detailed edit recognizer design for nine types of common face edits: faceswap, changing face expression, changing gender appearance, changing skin tone, changing hair color, adding/removing eyeglasses, changing age, changing face shape, and changing brightness of the photo. For each face edit type, we list (i) the corresponding attributor extractor that employs off-the-shelf models, and (ii) the *default* decision heuristics to detect the face edit.

We design one attributor extractor for each edit type. For the first eight face edit types, their attribute extractors are learning based, leveraging pre-trained deep learning classifiers to extract a person's age, identity vector, facial expression, gender appearance, skin color, hair color, face shape, and the presence of eyeglasses. In particular, to identify skin tone and hair color, we first apply a face segmentation model to locate the image pixels belonging to face and hair area, and then follow the human skintone color palettes (defined by [2]) and the hair color palettes (defined by [1]) to define the person's skin tone and hair color. Note that both palettes can be "reconfigured" by individual users. For eyeglasses, we first apply the face segmentation to identify the pixels belonging to eyeglasses area, then use them to determine whether any eyeglasses are present on the face. Finally, for the last edit type (changing brightness), we use a graphic metric **U** computed directly from the image pixel values. Here **U** is a common metric for calculating the average of (R+G+B) values across all the pixels in a photo, producing a value between 0 (dark) and 255 (bright).

Next, for each face edit type, we define the *default* decision rule, which can be reconfigured by a user's edit policy. In general, the default rule should be that the attribute extracted from the original photo (x_0) is different from that of the edited photo (x). For changing age, face shape and brightness, we include a value metric to be more specific. For age, we choose the difference to be more than 5 years; for face shape, we treat more than 5% changes in the detected face as an indicator of the edit; and for brightness, we empirically set a threshold of 10 to detect any "reasonable" brightness change. Furthermore, since some local edits on the image, e.g., changing hair color, could also change **U**, we add an additional requirement of more than half of pixels having brightness changes.

Key benefits. By comparing semantic attributes between x and x_0, Aletheia's edit recognizer achieves four key properties required for practical deployment:

- It is modular and scalable. Given a list of unacceptable edits, the system runs a set of stand-alone attribute extractors corresponding to these edits. As new face edits appear, the system can expand by adding new attribute extractors.
- It is tool-independent by identifying natural semantics of images rather than tool-specific features.
- It is agile against advancement of face edit tools, and remains effective even as edit tools perfect themselves to produce "natural" images without any artifacts.
- It is *reconfigurable*, allowing users to specify the type and amplitude of edits allowed or disallowed per attribute.

6.3 Prototype of Aletheia

We built an initial prototype of Aletheia in Python, leveraging existing libraries on *pHash* (a popular perceptual hash) and face attribute extractors. The prototype is configured to recognize 9 edit types (see Table 2) by employing public models as attribute extractors, and a default set of decision rules. We leave the design of a broader set of attribute extractors to future work. We used this prototype to evaluate the initial feasibility of §7. Our modular implementation provides extensibility – one can add new semantic attribute extractors or experiment with other image similarity metrics and image hashes for database search. We plan to release our code for academic use and expand the prototype to include more edit types.

7 EVALUATIONS

We evaluated Aletheia's effectiveness and usability using four forms of evaluations. All studies were approved by our institute's IRB. Using both user studies and experiments on large-scale datasets, we evaluated how Aletheia flags edited images that human users flagged as (un)acceptable (§7.1), performs on large image hosting services (§7.2) and addresses "in-the-wild" face edits (§7.3), and how users perceive Aletheia's protection on their online photos (§7.4). Later in §8 we also perform a security analysis on Aletheia against strongly motivated adversaries.

7.1 Aletheia's Decision vs. Human Decision

Using data from the user study in §3, we examined how Aletheia flags edited images that violate user policies, and whether such decisions match human decisions. For each participant, we used their responses to (1) define an edit policy per edit type, i.e., acceptable or unacceptable, and (2) obtain a set of human decisions on the edited images, which we use to evaluate Aletheia. The policy was generated from user data collected in step 1, and decision data from step 2. In total, we have 99 participants and 406 valid human decisions (156 unacceptable, 250 acceptable). Next, for each edited image labeled by humans, we ran Aletheia based on each participant's policy to determine whether Aletheia accurately detects those violating the policies. Our experiment produced two key findings.

Result #1: Aletheia can accurately flag edited images that users disallow (93.6%). We found that Aletheia's decisions match the participants' decisions well. It successfully flagged 93.6% of edited images (146 out of 156) that participants labeled as unacceptable, and accepted 87.6% of images (219 out of 250) that participants labeled as acceptable.

Result #2: Decision mismatch came from subtle edits and overlap of edits. We studied mismatch between Aletheia and participants' decisions, and found two dominating trends. First, the "unacceptable" images not detected by Aletheia *all* came down to a single skin tone edited image, which contains very subtle change of skin tone. Aletheia failed to spot the change because it uses a common human skin tone palette that "ignores" such subtle changes. Second, when Aletheia falsely flagged an acceptable image as unacceptable, the error came from overlap of edits. For example, an age edit often changes hair color, and face swap often changes expression, age, and face shape. When a participant's policy contains conflict across overlapped edits, e.g., allowing age edit but not hair color edit, those false alarms are inevitable.

Insight: the need for precise edit policy. Our results are encouraging and demonstrate an initial feasibility of Aletheia. They also confirm the observation that the current definition of edit types is likely too broad to build accurate edit recognizers. Aletheia could largely benefit from more precise characterization and interpretation of edit types, so users can clearly define fine-grained policies that are free of conflicts and can be implemented as decision rules.

7.2 Testing Aletheia at Scale

We also assess how Aletheia would perform on large image hosting services. Since a user study at this scale is intractable, we evaluated Aletheia on large-scale face datasets, exploring the accuracy of its image inspector and edit recognizer, and its computation cost.

Our face datasets. As no existing large-scale datasets provide edit type labels, we built our own dataset by altering original images with various editing tools. More details are in appendix A.

- *Original faces (820K images)*: Combining several public datasets, we built a diverse dataset of 820K face images across more than 30,400 identities.
- *Edited faces (42,500 images)*: We built scripts to generate edited images from 1000 randomly sampled face images, producing 42,500 edited images. Each image contains a single type of edit. For each edit type, we generate edited images using at least two tools. As shown in Table 11 (in appendix A), we used 10 different editing tools: 3 commercial tools (PhotoShop, PortraitPro, FaceApp) and 7 open-source tools (StarGAN, AttGAN, GANimation, HRFAE, OpenCV sticker code, FF++, DeeperForensics 1.0). We considered 12 edit types: 9 of them come from Table 2 for which Aletheia has built recognizers. We also included 3 extra edits (add filter, makeup, change eyenosemouth) for which Aletheia *does not* have recognizers designed. We used these three extra edits to evaluate false positives on Aletheia's 9 edit recognizers and the accuracy of Aletheia's image inspector.

Upload request x	Result of Image Inspector
x= new original face provenance=NULL	**99.54%: correctly identified as original** 0.46%: wrongly identified as edited
x= edited face provenance= x's original copy	**99.51%: correctly identified as edited, and paired with its original copy** 0.49%: correctly identified as edited, but paired with a wrong original copy
x= edited face provenance= NULL	**97.1%: correctly identified as edited and paired with its original copy** 2.9%: identified as original
x= edited face provenance = not x's original copy	**97.1%: correctly identified as edited, and paired with its original copy** 2.9%: correctly identified as edited, but paired with a wrong original copy

Table 3: ***The status of different upload requests after applying Aletheia's image inspector.***

Experiment configuration. Aletheia's performance depends on the configuration/scale of the database and the similarity threshold θ_{SSIM}. To ensure a fair evaluation, we split the original face dataset into 2 disjoint parts: 754,000 faces as the Aletheia's database of registered original faces, 43,000 faces that we will use to test Aletheia's image inspector. To set θ_{SSIM}, we randomly sampled 2,000 faces from the database to compute their SSIM scores, from which we set θ_{SSIM}=0.5 to reach a 1% false positive rate.

Result #3: Aletheia can accurately flag edited images and pair them with original versions (97.1%-99.5%). We tested Aletheia's image inspector using two datasets: the new original faces (43,000) and the edited faces (42,500). Each of these images is sent to Aletheia as an upload request. For the edited images, we also expanded the test set to consider three cases: the provenance data is either accurate, missing, or modified to change the declared original copy. Table 3 shows that Aletheia's image inspector identifies the edit status of these upload requests at high accuracy, i.e., 99.5% when the provenance data is intact and 97.1% when the provenance data is manipulated or removed.

Result #4: Aletheia largely outperforms today's face edit detectors. As detailed in Table 4, Aletheia's image inspector offers significant improvement over recent systems designed for generalized edit detection (FFD [13], CNNDetector [63]) and PhotoShop-specific edit detection (FAL [62]).

Result #5: Aletheia can recognize common face edits at a reasonable accuracy. We ran Aletheia's edit recognizer against the 42,500 edited face images paired with their original copies. Since Aletheia does not make any assumption on the number of edits in x, a single edit could trigger multiple edit recognizers, leading to false positives. We summarize the results in Table 5.

Today's Face Edit Detector	Accuracy (Original)	Accuracy (Edited)
FAL [62]	38.1%	37.0%
FFD [13]	41.8%	55.6%
CNNDetector [63]	93.5%	9.5%
Aletheia	**99.54%**	**97.1–99.51%**

Table 4: ***Aletheia significantly outperforms existing face edit detectors (FAL, FFD, CNNDetector) in terms of identifying whether an image is original or edited.***

High recognition rate (86.3% -99.4%): For the 9 edits that Aletheia attempts to recognize, the recognition rate is reasonably high. This is encouraging since our prototype just uses public, pre-trained models. The imperfect recognition rate is due to errors in attribute extractors (e.g., faceswap, gender, age) and/or imprecise decision rules (especially for color-related changes, e.g., haircolor, brightness). A more precisely designed user policy would help improve accuracy and match the diverse opinions of different users.

Moderate false positives: We observed visible false positives in the recognition result, due to natural overlap between edits and errors in attribute extractors. For example, since adjusting brightness also changes face color, some edit tools adjust skintone by adjusting brightness. Such overlap leads to 35.4% and 43.8% false positives between the two. Similarly, some edit tools apply aging by changing hair color, thus age edits often trigger the haircolor edit recognizer (26.5%). Finally, three edits (add filter, makeup, change EyeNoseMouth) not covered by Aletheia's recognizers also produced some false positives, again due to the overlap between edits.

Overall, these results match those in §7.1, demonstrating the initial feasibility of Aletheia and indicating the need for a more precise policy definition and interpretation.

Result #6: Aletheia is computationally efficient. We studied the end-to-end delay for Aletheia when running it on a server with a single CPU (Intel Xeon 2.2GHz) and single GPU (NVIDIA Titan RTX). When the input is an original image, the processing time is 939*ms* (all spent on the image inspector); for an edited photo, it is 924*ms* (24*ms* on the image inspector, 920*ms* on the edit recognizer). Note that these results were obtained on a low-end server rather than sophisticated servers used by photo hosting services.

7.3 Aletheia against In-the-Wild Face Edits

We also tested Aletheia on photos edited by real users, using their own tools that Aletheia has no knowledge of. We recruited 8 volunteers (non-authors), presented them with 100 face images (randomly chosen) and asked them to edit any of these as they wanted. The only instructions given were to log the edit(s) and make each edit visible by human eyes, with no restriction on the number of edits or what edit tools to use. We received 415 in-the-wild images, each with 1-4 different edits (1.9 average).

Result #7: Aletheia can flag unacceptable face edits done by real users. Of these 415 photos, Aletheia's image inspector correctly identified 391 (**94.2%**) as edited images and located their original copies. Next, Table 6 lists the recognition rate per edit type across these 391 images. The per-edit recognition rate is comparable to those in Table 5, except for gender appearance (97.2%) and skin tone (59%). We found that skin tone changes made by our volunteers were often subtle and thus "ignored" by Aletheia given its human skin color palette. While this can be largely mitigated by switching to a more fine-grained palette or applying a color-change threshold, it confirms the need for a more precisely defined policy matching each user's preferences.

7.4 User Perception of Aletheia's Protection

We conducted a *second* online survey to assess how users perceive the protection offered by Aletheia, and to submit, if any, suggestions on improving Aletheia. Our study was approved by the local Institutional Review Board (UChicago IRB-21-0502). Full survey script is available at https://sandlab.cs.uchicago.edu/faceedit/userstudy

Participants. We recruited 100 participants via Prolific. The survey was designed to take 10 minutes on average and participants received $2 as compensation. We received 97 valid responses (3 responses failed attention check question). Of those, 7 indicated they did not feel concerned at all about privacy online. Since those privacy-insensitive users are not our target users, we filtered their responses from our analysis. In the end, we analyzed 90 responses (44% identified as female, 66% male). The age distribution is: 18-29 years old (75%), 30-39 (16%), 40-49 (7%) and 50-59 (2%).

Task. We asked participants to imagine using Aletheia when posting an image online to a site like Instagram. We first show examples of each edit type, and demonstrate how the system would enforce potential policies when an unacceptably edited image is detected. We then asked multiple-choice and free response questions about the usability of the system, users' sense of protection, and their perceptions of privacy when posting images online.

This *conceptual approximation* of the Aletheia system captures its essence and demonstrates the potential value of the service. We used it to help our study participants understand the protection offered by Aletheia and determine whether they would want or need such protection. Also, since the remote/one-direction nature of our user study meant we could not debrief our remote participants, we chose to not collect or alter personal photos from our remote participants, in order to protect their privacy and minimize potential negative emotional effects.

Result #8: Many participants showed appreciation for the protection offered by Aletheia. We asked participants how they felt about the protection Aletheia would provide for their online photos. Table 7 shows a summary of the responses. Nearly half (48%) of the participants felt that Aletheia protected their images, especially since they can define personalized protection policy. 15% of the participants were neutral. They questioned the full effectiveness of the protection, but still viewed Aletheia as a step in the right direction. 13.3% of the participants did not feel protected by Aletheia because they worried that the system could be bypassed, such as posting edited images elsewhere online, or were not convinced Aletheia's technology could accurately detect most edits. 23.7% of the participants expressed that posting images online is never safe and the only way of protection is not uploading any.

Result #9: Many participants would like to use Aletheia. Regarding whether they would use Aletheia to protect their online images, we observed considerable differences between *edit-concerned* and *edit-unconcerned* participants (see Table 8). Note that at the begining of the user survey, we asked each participant whether

Edit(s) recognized by Aletheia	Edit in the Image											
	Brightness	Skintone	Haircolor	Faceswap	Gender	Age	Faceshape	Expression	Eyeglasses	Filter	Makeup	EyeNoseMouth
Brightness	**89.6%**	43.8%	4.9%	0.0%	0.4%	0.3%	0.0%	0.0%	0.0%	27.5%	0.0%	0.0%
Skintone	35.4%	**99.4%**	12.3%	3.5%	11.8%	9.9%	1.4%	2.3%	4.6%	40.2%	4.3%	0.9%
Haircolor	4.3 %	5.1 %	**88.0%**	1.6%	38.1%	26.5%	1.6%	1.0%	2.1%	30.5%	1.0%	0.8%
Faceswap	0.2%	0.0%	12.0%	**86.3%**	39.4%	10.2%	0.0%	0.0%	1.2%	0.0%	0.0%	0.0%
Gender	3.5%	5.0%	5.6%	7.5%	**88.8%**	9.5%	6.0%	8.7%	10.3%	3.6%	12.2%	3.4%
Age	7.8%	20.5%	33.6%	62.3%	42.2%	**88.5%**	9.8%	21.6%	30.9%	11.1%	8.0%	2.2%
Faceshape	2.4%	9.0%	23.9%	5.6%	36.6%	16.2%	**97.0%**	1.2%	9.6%	3.6%	0.9%	0.6%
Expression	6.8%	8.7%	9.6%	17.6%	14.0%	11.5%	7.1%	**95.4%**	12.3%	4.5%	6.4%	3.5%
EyeGlasses	1.0 %	2.7 %	3.8%	4.6%	5.2%	3.2%	4.7%	1.8%	**91.8%**	1.1 %	0.8%	0.4%

Table 5: ***Results on how each edit on an image gets recognized by Aletheia. Each column refers to a specific type of edit e contained by the image, where the bold entry in the column is the probability the edit e is recognized by Aletheia, and the other entries are the false positives on other 8 recognizers triggered by e.***

Brightness	Skin tone	Faceswap	Expression	Age
75.8%	59.1%	75.0%	82.6%	81.2%

Hair color	Faceshape	Eyeglasses	Gender appearance
80.8%	77.6%	94.1%	97.2%

Table 6: ***The recognition rate of Aletheia's edit recognizers on in-the-wild face edits.***

Response	Reason	Example
Protected (48%)	Trust system works to detect disallowed edits	"I would feel my images **are protected by the system as I can specify** whether I would like them to be modifed [*sic*] in a way I would not like."
Neutral (15%)	Can't 100% guarantee protection	"**it may miss when a photo has been edited**" "i think they protect the images **to a certain extent however not fully**"
Not Protected (13.3%)	The system can be bypassed (8.5%)	"I think it **could be cheated** easilly [*sic*]" "Pictures can still be extracted and **posted somewhere else**."
	Don't trust system (4.8%)	"I **don't think the system is advanced enought** [*sic*] to detect these images."
Never (23.7%)	Posting images online is never safe	"I think it's **never safe when we post pictures of ourselves** because they never really leave the internet."

Table 7: ***Participant responses for whether they feel their images were protected with Aletheia.***

User group	Yes	Neutral	No
edit-concerned	68%	21%	11%
edit-unconcerned	42%	27%	27%

Table 8: ***Participant responses for whether they would use Aletheia when posting images on social media sites.***

they are concerned about their image being edited and reposted by others, and the result was a near-even split (49%/51%) across participants. From Table 8, we see that 68% of *edit-concerned* participants were interested in using Aletheia and 21% were neutral. Of the 11% (5 participants) who said no, 4 expressed that they **never** shared images on social platforms and did not feel protected even with Aletheia. Another interesting observation is that even among those not concerned with edit, 42% indicate they would use Aletheia.

Overall, our survey results are highly encouraging, showing that most participants express interest in using Aletheia to increase protection online. More efforts like Aletheia should be made to provide more privacy-friendly services and to educate users on ways to achieve their privacy goals.

Result #10: Participants want to configure and adapt their edit policy, despite the overhead. We surveyed how participants feel about configuring their face edit policy, and how their policy may change over time.

First, we asked how users consider the tradeoff between time spent setting up their own policy, and achieving protection. Most participants were either not concerned (45%) or neutral (32%), deeming the protection worth the initial setup time. The rest 23% expressed concern about the time spent, with one participant feeling this may leave many users reverting to default settings. Also, 75% of participants indicated they would prefer a single policy for all images, for simplicity and efficiency. Second, similar to the first study, we found that participants want flexibility to change their preferences over time, and expect the system to adapt and new editing methods are developed. Together, these feedbacks suggest that the design of Aletheia's edit policy management should serve to spare the users' efforts, whilst affording personalized control.

Suggestions on improving Aletheia. We asked participants what changes, if any, they would make to improve Aletheia. While most participants did not submit any response, there are a few notable ones. 4 participants indicated they would like notifications for *any* edits detected, so they could decide whether to remove them. 11 participants care about *who* makes the edit, such as "set certain friends to have edit permissions" or "allow users to ask for permissions to the original author of the image." Finally, several participants brought up a desire to implement Aletheia on all possible platforms, providing ultimate protection against any edited face images posted online. We agree that this is a natural follow-up work and discuss it next in §9.

8 ROBUSTNESS UNDER STRONG ATTACKS

As mentioned in §5, our system is designed under the threat model of face edits made by users familiar with everyday technology, rather than highly skilled and resourceful adversaries. However, it is important to also consider the potential effects of more powerful threats. In this section, we investigate the robustness of our current design against strong attackers with security expertise and significant computational resources, and attacks that degrade the visual quality of the image, as well as possible defenses.

8.1 Attacking Aletheia's Image Inspector

For each uploaded image, Aletheia first determines whether it is new or an edited copy (of a photo in its database). An attacker can alter an edited image x to evade Aletheia's hash-based image search, such that the image inspector either misclassifies it as an original image, or associates it with an incorrect original photo (one that has relaxed edit restrictions). Here we consider attacks that introduce significant modifications to the image (rotating or cropping the face), and those that apply complex optimization to generate pixel-level perturbations that distort the pHash values.

Note that we already discussed standard attacks such as deleting or modifying metadata (§6.1), and shows (Table 3) that Aletheia already resists this attack by applying verification via hash-based image search and SSIM comparison (i.e., Step2a and 2b).

pHash evasion using significant image modifications. Recent studies have shown that pHash-based image search could be misled by applying image post-processing [22]. We examined the impact on Aletheia using three common post-processing techniques: scaling, rotation, and cropping. We randomly selected 1000 images (x) from our *edited* image dataset and located their original copy (x_0). We applied each of the three processing techniques on these edited images (x), producing their processed copy (x^P). We then ran Aletheia's image inspector on these processed images (x^P) and examined whether Aletheia can still detect x^P as edited and locate their original copy (x_0). For a fair evaluation, we set their provenance to empty. Our results (described below) show that Aletheia's pHash-based image search is insensitive to scaling, but is less robust against rotation and cropping.

- Scaling: when we apply scaling to an edited image x, ranging from a factor between 50% and 200%, Aletheia correctly identified them as edited and located the true original copy x_0 associated with all scaled images. This is because pHash computation normalizes image size to a 32×32 pixel matrix, which nullifies the impact of scaling.
- Cropping: when we crop each edited image to remove 4%, 8%, 12% and 16% of the face content, the probability of detecting them as edited and locating the true original copy reduces from 100% (no cropping) to 99.8%, 73.8%, 14.9% and 2%, respectively. The result is also shown in the first row (w/o aug.) in Table 9.
- Rotation: When we rotate edited images by 2°, 3°, 5° and 7°, detection probability drops to 88%, 52.2%, 5.3% and 0.4%.

To address the impact of rotation and cropping, one potential solution is to register multiple rotated and/or cropped versions of an original image during registration for Aletheia's database. In our experiments, we found that augmenting each image in Aletheia's database with two extra versions ("removed by 12%" and "rotated by 5°"), the detection accuracy improves considerably (as shown by the row of "w/ aug" in Table 9). In this defense, each original image has three (rather than one) hash values. This also means that Aletheia does not need to make an extra hash version per modification instance, i.e., the defense is scalable. Another potential protection against rotation is to normalize the rotation of the face in all photos, *i.e.* rotating the face to a strictly front-facing position [57]. This, however, increases computation overhead.

	Cropping				Rotating			
	4%	8%	12%	16%	2°	3°	5°	7°
w/o aug.	99.8%	73.8%	14.9%	2%	88%	52.2%	5.3%	0.4%
w/ aug.	99.8%	90%	100%	95.7%	88%	84%	100%	95%

Table 9: ***Aletheia's detection accuracy largely improves after augmenting the database with two cropped and rotated versions of the original photos.***

pHash evasion using pixel-level perturbations. Finally, a more capable attacker can run complex optimization to compute pixel-level perturbation on an edited image x to enlarge the hash distance to the unperturbed x while minimizing visual changes [22]. The optimization can either try to make the perturbed x's pHash significantly different from all original photos in Aletheia's database, so Aletheia misidentifies it as original; or it can make the pHash similar to another original image x_0' that has no (or weaker) edit restrictions, so Aletheia misidentifies x as an edited copy of x_0' and admits it. While our threat model does not assume this type of technically advanced attackers, Aletheia is vulnerable to this type of pHash evasion attack.

One promising defense against such attacks, described in [22], is to apply pre-processing (e.g., blurring) to photos before pHash computation to reduce the impact of potential perturbations. Another (complementary) defense is to add a step of face identity verification after Aletheia pairs an edited image with its original version. Assuming the attacker has no control of the user's images and their edit policies, the chosen x_0' in the above attack will have an identity different from that of x. Finding pixel perturbation that misleads both the pHash-based image search and the face recognition is a very difficult challenge.

Overall, we expect researchers to continue to develop increasingly powerful attacks and defenses for hash-based image search [22]. While the cat-and-mouse game will likely continue, we hope advanced defenses will raise the bar for successful attacks well above the level expected from our everyday user threat model.

8.2 Attacks against Aletheia's Edit Recognizer

Motivated and resourceful adversaries can also target Aletheia's edit recognizer to disguise a forbidden edit as an allowed one. Specifically, an attacker can carefully craft the edit so that the corresponding face attribute extractor employed by Aletheia will produce an inaccurate result that prevents the edit from being detected or exceeding the allowed range. For example, the attacker first edits a 20 years old's face photo to make them 40 years old, then adds carefully computed adversarial pixel perturbations on the photo so that Aletheia's age classifier misclassifies the edited photo as 20 years old. Thus the age change is not detected.

White-box evasion attacks against attribute extractors. To launch these attacks, the attacker generally needs white-box access to the deep learning models used by Aletheia, i.e., the attacker has total access to the target model, including its internal architecture, weights and parameters. With this, an advanced attacker with sufficient compute resources can generate the required pixel perturbations for the current photo as an optimization problem. There are already numerous defense proposals and ongoing works that seek to either prevent the generation of adversarial perturbations or detect them at run-time (e.g., [42, 55, 61, 66]). We expect

any practical deployment of Aletheia to leverage these existing and ongoing efforts, and adopt attribute extractors that are more robust against such attacks.

Black-box evasion attacks against attribute extractors. A highly advanced and determined attacker can also apply black-box query-based attacks against Aletheia's face attribute extractors. Here the attacker does not have access to the model, but conducts repeated queries to Aletheia and adapts the pixel perturbations in the edited image until it gets admitted by Aletheia. Fortunately, these attacks require thousands to hundreds of thousands of queries to produce a successful attack. In practical settings, an image sharing platform can easily detect and flag a high volume of rejected photo uploads. They can also leverage more advanced defenses that detect black-box query-based attacks in the image domain [32].

As future work, we plan to integrate Aletheia with both robust attribute extractors and defenses against query-based attacks [32], and conduct more in-depth studies on robustness against these adversarial attacks. Again, our goal is to raise the attack cost well above the level expected from our everyday user threat model.

9 LIMITATIONS AND FUTURE WORK

As the first work on face edit protection for online photos, Aletheia faces a number of limitations, much of which will be the targets of future work in this space. Beyond addressing stronger attackers (§8), we outline below additional directions for future work.

(1) Deeper study on users' tolerance for face edits: Our user study is limited in that we collected tolerance of different face edits when participants evaluated others' face photos (to protect our participants). This tolerance may change when participants evaluate their own individual photos, which needs to be considered when collecting the specific edit policy from a user seeking protection.

(2) Edit policy definition and management: Our current edit policy specification adopts a simple (default) policy on several common types of face edits. We recognize three broad challenges in clearly defining and deploying face edit policies.

- Current tools and literature define broad and vague "types" of face-edits, and many edits are naturally correlated. These have affected the accuracy of Aletheia's edit recognition. We need a systematic approach to interpret and decompose face edit types, and an interactive interface to guide users in defining usable policy. Here a related question is how to effectively illustrate the edit effect to users while minimizing/addressing potential negative emotional impacts.
- Defining certain edit types such as gender appearance and age may rely on common stereotypes that fail to properly capture real world diversity. Much work remains in developing a more nuanced and powerful policy specification that better reflects user diversity.
- The third challenge is how to automate policy configuration. One can explore the use of machine learning tools to learn users' preferences, and help them set their edit policy automatically.

(3) Expanding edit recognizer: So far our prototype employs nine attributor extractors built from public models. We plan to add new ones to cover a broader range of edits, leveraging ongoing efforts on semantic face analysis. This effort needs to be integrated with the policy component to meet the needs of real-world users.

(4) Integration with multiple photo-sharing platforms: So far Aletheia targets a single photo-sharing platform. While this can be effective to protect users if deployed by a very large platform like Instagram, we could achieve much more impact if multiple platforms collaborate. Thus a natural extension to this work would consider privacy-preserving ways to share personalized user policies and data across platforms, so that unacceptable edits of images from one platform can be detected on others.

(5) Addressing detection errors: Like any practical system, Aletheia may occasionally make mistakes. Here we discuss two main types of errors and ways to mitigate them. The first type is wrongly recognizing a new image as an edited one[3] and forwarding it to the wrong owner to review. One way to reduce the likelihood of such errors is to add a verification step to check whether the face identities of the edited image and its original copy match, *i.e.* the two images are photos of the same person. When the two images display different identities, it could be a detection error or caused by a "faceswap" edit. Such cases could be reviewed by the platform's moderator before taking further actions.

The second type of errors is wrongly identifying an edited image as original, or failing to detect the disallowed edits, so the image is posted online. A user affected by this type of error can mark the photo and submit a complaint. Aletheia can verify the complaint and remove the image post if necessary. In addition, Aletheia can use this data point to diagnose and improve its detection algorithms. Thus real-world deployments of Aletheia need to include a mechanism for users to report errors.

(6) Verifying photo ownership: Aletheia protects each original photo based on its edit policy. Intuitively, the legal owner(s) of an original photo should be the one who defines the policy. This leads to the issue of how to define the legal owner(s) of a photo [43], e.g., the person who took the photo, or the person who owns the copyright to the photo. This ownership issue should be addressed by each photo-sharing platform before deploying Aletheia, e.g., via their term-of-service or copyright agreement.

10 CONCLUSION

Our work seeks to address the threat of online face photos getting edited and reposted by others for malicious purposes. Our user study shows that users are concerned about this threat and want actions taken to protect their online photos. But realizing such protection is challenging because users vary widely in their definition of what edits are (un)acceptable. This motivates us to develop an image moderation tool that online platforms can deploy to provide personalized protection against unacceptable face edits. In this work, we design and prototype Aletheia to address two immediate challenges of personalized face edit protection: detecting and recognizing individual edits on a photo and also identifying its original version (and thus its edit policy).

Overall, our work demonstrates the initial feasibility for online platforms to support social interactions via photo editing and sharing, while giving users agency over how their photos can be altered by others. To the best of our knowledge, our work is the first to explore and propose solution to this real-world problem. We hope it spurs more efforts to reduce potential misuse of face editing.

[3]Results in Table 3 show that this is of very low probability, 0.46%.

ACKNOWLEDGMENTS

We thank our anonymous reviewers for their insightful feedback. We also thank Marshini Chetty for her feedback on the paper. Zhujun Xiao, Jenna Cryan, Yuanshun Yao, Yi Hong Gordon Cheo, Ben Y. Zhao and Haitao Zheng were supported in part by NSF grants CNS-1949650 and CNS-1923778. Opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of any funding agencies.

REFERENCES

[1] Hair color palette. https://colorswall.com/palette/43520/.
[2] Skin tone palette. https://www.summitprintingpro.com/graphic-design/tutorials/skin-tone-correction.html.
[3] Adobe. Content authenticity initiative (cai). https://contentauthenticity.org/approach, 2020.
[4] Agarwal, S., Farid, H., El-Gaaly, T., and Lim, S.-N. Detecting deep-fake videos from appearance and behavior. *arXiv:2004.14491* (2020).
[5] Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., and Li, H. Protecting world leaders against deep fakes. In *Proc. of CVPR workshops* (2019).
[6] Anderson, M. A majority of teens have experienced some form of cyberbullying. https://www.pewresearch.org/internet/2018/09/27/a-majority-of-teens-have-experienced-some-form-of-cyberbullying/, Sept. 2018.
[7] Arriaga, O., Valdenegro-Toro, M., and Plöger, P. Real-time convolutional neural networks for emotion and gender classification. *arXiv:1710.07557* (2017).
[8] Besmer, A., and Richter Lipford, H. Moving beyond untagging: photo privacy in a tagged world. In *Proc. of CHI* (2010).
[9] Blackwell, L., Dimond, J., Schoenebeck, S., and Lampe, C. Classification and its consequences for online harassment: Design insights from heartmob. *Proc. of CSCW* (2017).
[10] Blackwell, L., Ellison, N., Elliott-Deflo, N., and Schwartz, R. Harassment in social virtual reality:challenges for platform governance. *Proc of CSCW* (2019).
[11] Buchner, J. A python perceptual image hashing module. https://github.com/JohannesBuchner/imagehash, 2021.
[12] Choi, T. R., and Sung, Y. Instagram versus snapchat: Self-expression and privacy concern on social media. *Telematics and Informatics 35*, 8 (2018).
[13] Dang, H., Liu, F., Stehouwer, J., Liu, X., and Jain, A. K. On the detection of digital face manipulation. In *Proc. of CVPR* (2020).
[14] Dhir, A., Torsheim, T., Pallesen, S., and Andreassen, C. S. Do online privacy concerns predict selfie behavior among adolescents, young adults and adults? *Frontiers in Psychology 8* (2017).
[15] Donghyeon, W. Gender and race classification with face images. https://github.com/wondonghyeon/face-classification.
[16] Dufour, N., et al. Deepfakes detection dataset by Google & Jigsaw, 2019.
[17] Egelman, S., Oates, A., and Krishnamurthi, S. Oops, i did it again: Mitigating repeated access control errors on facebook. In *Proc. of CHI* (2011).
[18] Fan, J., and Zhang, A. X. Digital juries: A civics-oriented approach to platform governance. In *Proc. of CHI* (2020).
[19] Gill, R. Changing the perfect picture: Smartphones, social media and appearance pressures, March 2021. https://www.city.ac.uk/__data/assets/pdf_file/0005/597209/Parliament-Report-web.pdf.
[20] Gordon, S. Why kids are using Instagram to bully. VeryWellFamily, December 2019. https://www.verywellfamily.com/how-kids-use-instagram-to-bully-460579.
[21] Güera, D., and Delp, E. J. Deepfake video detection using recurrent neural networks. In *Proc. of International Conference on Advanced Video and Signal Based Surveillance (AVSS)* (2018).
[22] Hao, Q., Luo, L., Jan, S. T., and Wang, G. It's not what it looks like: Manipulating perceptual hashing based applications. In *Proc. of SIGSAC* (2021).
[23] Hasan, R., Li, Y., Hassan, E., Caine, K., Crandall, D. J., Hoyle, R., and Kapadia, A. Can privacy be satisfying? on improving viewer satisfaction for privacy-enhanced photos using aesthetic transforms. In *Proc. of CHI* (2019).
[24] Hohman, M. Cheerleader's mom accused of using deepfakes to harass girl on team, March 2021. https://www.today.com/news/cheerleader-s-mom-accused-using-deepfakes-harass-girl-team-t211737/.
[25] Huh, M., Liu, A., Owens, A., and Efros, A. A. Fighting fake news: Image splice detection via learned self-consistency. In *Proc. of ECCV* (2018).
[26] Iacobucci, S., De Cicco, R., Michetti, F., Palumbo, R., and Pagliaro, S. Deepfakes unmasked: The effects of information priming and bullshit receptivity on deepfake recognition and sharing intention. *Cyberpsychology, Behavior, and Social Networking 24* (2021).
[27] Jiang, L., Li, R., Wu, W., Qian, C., and Loy, C. C. Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proc. of CVPR* (2020).
[28] Karras, T., Laine, S., and Aila, T. A style-based generator architecture for generative adversarial networks. In *Proc. of CVPR* (2019).
[29] Knight, W. Facebook is making its own AI deepfakes to head off a disinformation disaster. MIT Tech. Review, Sept. 2019.
[30] Korus, P. Digital image integrity–a survey of protection and verification techniques. *Digital Signal Proc. 71* (2017).
[31] Krause, A. People are editing photos of celebrities to give them Instagram-inspired faces. experts say it could be harmful, 2020. https://www.insider.com/why-edited-photos-of-celebrities-can-be-harmful-2020-9.
[32] Li, H., Shan, S., Wenger, E., Zhang, J., Zheng, H., and Zhao, B. Y. Blacklight: Scalable defense for neural networks against Query-Based Black-Box attacks. In *Proc. of USENIX Security* (2022).
[33] Li, Y., Chang, M.-C., and Lyu, S. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *Proc. of International Workshop on Information Forensics and Security (WIFS)* (2018).
[34] Li, Y., and Lyu, S. Exposing deepfake videos by detecting face warping artifacts. In *Proc. of CVPR Workshops* (2019).
[35] Li, Y., Vishwamitra, N., Hu, H., and Caine, K. Towards a taxonomy of content sensitivity and sharing preferences for photos. In *Proc. of CHI* (2020).
[36] Li, Y., Vishwamitra, N., Knijnenburg, B. P., Hu, H., and Caine, K. Effectiveness and users' experience of obfuscation as a privacy-enhancing technology for sharing photos. *Proc. of CSCW* (2018).
[37] Liu, T., Moore, A. W., and Gray, A. New algorithms for efficient high-dimensional nonparametric classification. *Journal of Machine Learning Research 7* (2006).
[38] Liu, Y., Nakatsuka, Y., Sani, A. A., Agarwal, S., and Tsudik, G. Videoprov: Verifiable provenance for videos from mobile devices. In *Proc. of MobiSys* (2022).
[39] Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proc. of ICCV* (2015).
[40] London, L. How beauty filters are making us 'look better' but feel worse. https://www.forbes.com/sites/lelalondon/2020/03/23/in-self-isolation-filter-dysmorphia-and-beauty-filters-will-threaten-our-mental-health/?sh=370b0c903831, 2020.
[41] Lorenz, T. Teens are being bullied 'constantly' on instagram. The Atlantic, Oct 2018. https://www.theatlantic.com/technology/archive/2018/10/teens-face-relentless-bullying-instagram/572164/.
[42] Madry, A., et al. Towards deep learning models resistant to adversarial attacks. *arXiv:1706.06083* (2017).
[43] Marshall, C. C., and Shipman, F. M. Who owns the social web? *Commun. ACM 60*, 5 (apr 2017), 52–61.
[44] Mlot, S. Instagram launches new features to curb online bullying. PC Magazine, May 2020. https://www.pcmag.com/news/instagram-launches-new-features-to-curb-online-bullying.
[45] Mondal, M., Yilmaz, G. S., Hirsch, N., Khan, M. T., Tang, M., Tran, C., Kanich, C., Ur, B., and Zheleva, E. Moving beyond set-it-and-forget-it privacy settings on social media. In *Proc. of CCS* (2019).
[46] NVlabs. NVlabs FFHQ Dataset. https://github.com/NVlabs/ffhq-dataset, 2019.
[47] Quinn, K., Epstein, D., and Moon, B. We care about different things: Non-elite conceptualizations of social media privacy. *Social Media+ Society 5*, 3 (2019).
[48] Rood, M. L., and Schriner, J. The internet never forgets: Image-based sexual abuse and the workplace. *Handbook of Research on Cyberbullying and Online Harassment in the Workplace* (2021), 107–128.
[49] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Niessner, M. FaceForensics++: Learning to detect manipulated facial images. In *Proc. of ICCV* (2019).
[50] Rothe, R., Timofte, R., and Van Gool, L. Dex: Deep expectation of apparent age from a single image. In *Proc. of ICCV workshops* (2015).
[51] Rothe, R., Timofte, R., and Van Gool, L. Deep expectation of real and apparent age from a single image without facial landmarks. *International Journal of Computer Vision 126* (2018).
[52] Ruiz, N., Bargal, S. A., and Sclaroff, S. Disrupting deepfakes: Adversarial attacks against conditional image translation networks and facial manipulation systems. In *European Conference on Computer Vision* (2020), Springer, pp. 236–251.
[53] Ryan-Mosley, T. Beauty filters are changing the way young girls see themselves, 2021. https://www.technologyreview.com/2021/04/02/1021635/beauty-filters-young-girls-augmented-reality-social-media/.
[54] Schroff, F., Kalenichenko, D., and Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *Proc. of CVPR* (2015).
[55] Shan, S., Wenger, E., Wang, B., Li, B., Zheng, H., and Zhao, B. Y. Gotta catch 'em all: Using honeypots to catch adversarial attacks on neural networks. In *Proc. of CCS* (2020).
[56] Squicciarini, A. C., Sundareswaran, S., Lin, D., and Wede, J. A3p: adaptive policy prediction for shared images over popular content sharing sites. In *Proc. of ACM conference on Hypertext and hypermedia* (2011).
[57] Steinebach, M., Berwanger, T., and Liu, H. Towards image hashing robust against cropping and rotation. In *Proceedings of the 17th International Conference on Availability, Reliability and Security* (2022).

Sub dataset	# of identities	# of images	Type (Source)
CelebA [39]	10,177	202,599	celebrities (Internet)
FFHQ [28]	unknown	70,000	normal people (Flickr)
DeeperForensics [27]	unknown	1,000	faces (YouTube video, using the first frame)
FF++ [49]	unknown	1,000	faces (YouTube video, using the first frame)
IMDB-WIKI [50]	20,284	523,051	actors (IMDb, Wikipedia)
UTKFace [67]	unknown	23,708	faces (Internet)
Total	**>30,461**	**821,358**	**normal people & celebrities**

Table 10: Our original face dataset includes 820K+ face photos from both normal people and celebrities.

Category	Edit type	# of images	Edit tools
Global processing	Add filter	3,941	FaceApp, PortraitPro
	Change brightness	5,936	PhotoShop, PortraitPro
Modify facial attributes	Change age	4,059	FaceApp, StarGAN, HRFAE
	Change gender appearance	2,000	AttGAN, StarGAN
	Change face shape	2,954	PhotoShop, PortraitPro
	Change skin tone	2,376	AttGAN, PortraitPro
	Change hair color	2,486	FaceApp, StarGAN
	Resize eye/nose/mouth	7,914	PhotoShop, FaceAPP, PortraitPro
	Add makeup	4,925	FaceApp, PortraitPro
Change facial expression		1,967	FaceAPP, GANimation
Add/Remove Eyeglasses		1,989	FaceApp, OpenCV code
Change face identity (facewap)		2,000	FF++, DeeperForensics
12 Edit Types		**42,547**	**10 Edit Tools**

Table 11: Edited faces: we generated and labeled more than 42K edited images, covering 12 popular face editing types and 10 popular edit tools (3 commercial and 7 open-source tools).

[58] Such, J. M., Porter, J., Preibusch, S., and Joinson, A. Photo privacy conflicts in social media: A large-scale empirical study. In *Proc. of CHI* (2017).

[59] Technologies, W. Stop cyber-bullying in its tracks. https://contentauthenticity.org/approach, 2011.

[60] Tiggemann, M., Anderberg, I., and Brown, Z. Uploading your best self: Selfie editing and body dissatisfaction. *Body image 33* (2020).

[61] Tramèr, F., et al. Ensemble adversarial training: Attacks and defenses. In *Proc. of ICLR* (2018).

[62] Wang, S.-Y., Wang, O., Owens, A., Zhang, R., and Efros, A. A. Detecting photoshopped faces by scripting photoshop. In *Proc. of ICCV* (2019).

[63] Wang, S.-Y., Wang, O., Zhang, R., Owens, A., and Efros, A. A. CNN-generated images are surprisingly easy to spot... for now. In *Proc. of CVPR* (2020).

[64] Yang, X., Li, Y., and Lyu, S. Exposing deep fakes using inconsistent head poses. In *Proc. of ICASSP* (2019).

[65] Yu, N., Davis, L. S., and Fritz, M. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proc. of CVPR* (2019).

[66] Zantedeschi, V., Nicolae, M.-I., and Rawat, A. Efficient defenses against adversarial attacks. In *Proc. of AISec* (2017).

[67] Zhang, Z., Song, Y., and Qi, H. Age progression/regression by conditional adversarial autoencoder. In *Proc. of CVPR* (2017).

[68] Zhou Wang, Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing 13*, 4 (2004).

[69] zllrunning. Face segmentation model. https://github.com/zllrunning/face-parsing.PyTorch, 2019.

A DETAILED DESCRIPTION OF OUR FACE DATASETS

Evaluation of Aletheia requires a dataset covering a wide range of face edit types and tools. As no existing datasets provides edit type labels, we built our own dataset by altering real face images with editing tools using automatic scripts.

- *Original faces* – By combining several public datasets (see Table 10), we built a dataset containing 821,358 face images of >30,461 identities. This combined set ensures diversity and includes a wide variety of images from celebrities and normal people. For consistency, each image only contains a single face.

Category	Example face edit types
Global retouch	change photo brightness; add filter effect
Insert sticker	add sunglasses/emoji
Change facial attributes	increase/decrease age, change gender appearance, add/remove hair, change face shape; add makeup
Change expression	non-smile → smile, smile → crying
Change identity	swap two faces

Table 12: *The face edit types considered by our study.*

- *Edited faces* – We built and ran scripts to generate edited images from 1000 original face images (randomly sampled), producing 42.5K edited images labeled by the edits. As detailed in Table 11, our dataset covers 12 edit types and 10 editing tools, including both commercial software/apps (PhotoShop, PortraitPro, FaceApp) and open-source models (StarGAN, AttGAN, GANimation, HRFAE, OpenCV sticker code, FF++, DeeperForensics 1.0). Each image contains a single type of edit. For each edit type, we generate edited images using at least two different tools. Due to variations in both the number of available tools and their edit options, our edited face dataset is not balanced across edit types. To avoid bias, we up-sampled under-represented types when reporting results that aggregate over edit types.

B USER STUDY DETAILS

Here we show the context provided to the participants, and the interface of the survey with examples. The full scripts for both user studies can be found at https://sandlab.cs.uchicago.edu/faceedit/userstudy.

Context Establishment. Suppose you're sharing a photo online to a platform, similar to Facebook or Instagram. Similar to when you post pictures to these online platforms, other people who can view the picture may edit your photo and upload it to the same platform. Some people may do this for fun (e.g., add fun stickers). Other people may do this maliciously (e.g., cyberbullying).

This platform detects if an image has been edited and re-uploaded. When you upload an image, you can specify a set of preferences associated with the image. Each preference setting either allows or disallows a particular type of editing. After an image is uploaded, the platform can detect and remove any of your pictures that have been edited in a way that violates your current settings.

Figures 5, 6, 7, 8 show examples of the survey interface.

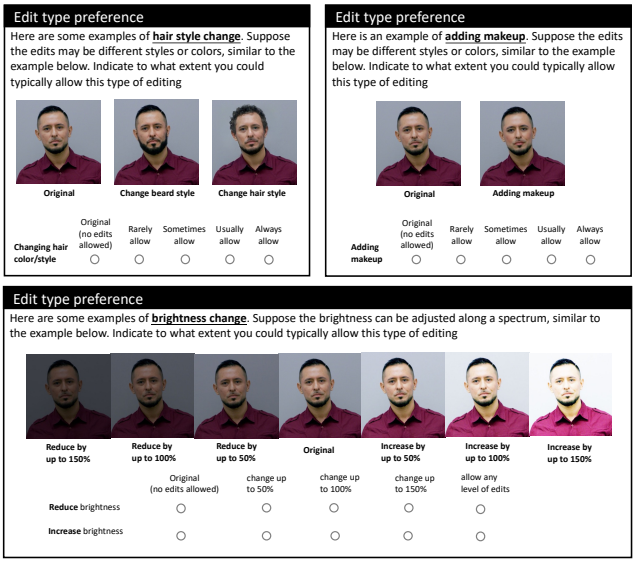

Figure 5: ***Example survey questions in our user study. We ask participants to rate their tolerance of different face edit types.***

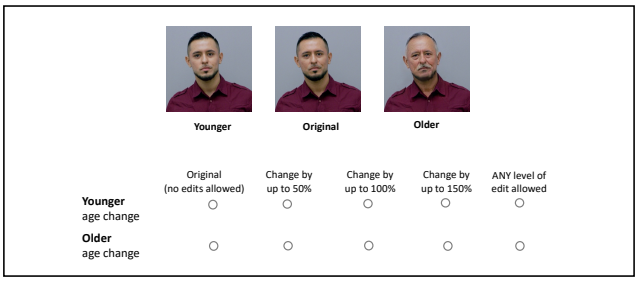

Figure 6: Question about preferences for changing age.

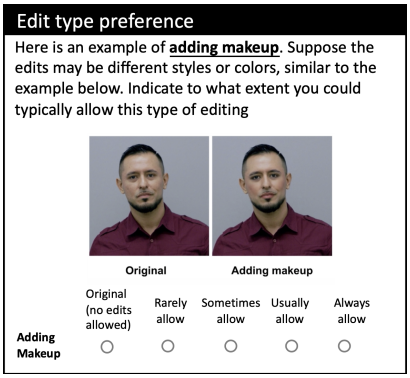

Figure 7: Example question shown to participants to illustrate how users of Aletheia specify their edit preferences.

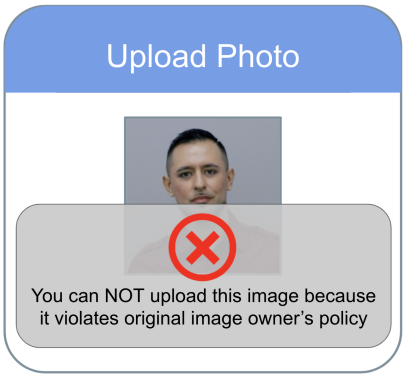

Figure 8: Aletheia blocks the image upload because it violates the original image's policy.