Development of two novel face-recognition CAPTCHAs: a security and usability study

Guido Schryen[a], Gerit Wagner[a,*], Alexander Schlegel[a]

[a] *Department of Management Information Systems, Universität Regensburg, Universitätsstr. 31, 93053 Regensburg, Germany*

Abstract

CAPTCHAs are challenge-response tests that aim at preventing unwanted machines, including bots, from accessing web services while providing easy access for humans. Recent advances in artificial-intelligence based attacks show that the level of security provided by many state-of-the-art text-based CAPTCHAs is declining. At the same time, techniques for distorting and obscuring the text, which are used to maintain the level of security, make text-based CAPTCHAs difficult to solve for humans, and thereby further degrade usability. The need for developing alternative types of CAPTCHAs which improve both, the current security and usability levels, has been emphasized by several researchers.

With this study, we contribute to research through (1) the development of two new face recognition CAPTCHAs (*Farett-Gender* and *Farett-Gender&Age*), (2) the security analysis of both procedures, and (3) the provision of empirical evidence that one of the suggested CAPTCHAs (Farett-Gender) is similar to Google's reCAPTCHA and better than KCAPTCHA concerning effectiveness (error rates), superior to both regarding learnability and satisfaction but not efficiency.

Keywords: CAPTCHA, Usability, Facial features, Gender classification, Age classification, Face recognition reverse Turing test

[*] Corresponding author. Tel.: +49 941 943 5743; fax: +49 941 943 5635.

Email addresses: guido.schryen@ur.de (Guido Schryen), gerit.wagner@ur.de (Gerit Wagner), alexander.schlegel@stud.uni-regensburg.de (Alexander Schlegel)

1. Introduction

CAPTCHA[1] (**C**ompletely **A**utomated **P**ublic **T**uring test to tell **C**omputers and **H**umans **A**part) procedures [7, 44, 89, 90, 92] are challenge-response tests that aim at preventing unwanted machines, including bots, from commenting spam in blogs, registering at websites, harvesting email addresses and conducting dictionary attacks, among others [16]. The need for protecting web services from unauthorized access is high, as several incidents have shown [64, 57, 62, 73].

Captchas are designed to be easy for humans but difficult for computers. Several big companies have suggested and applied Captchas [14], including Google's *reCAPTCHA* [93], *eBay Captcha*, *Yahoo Mail Captcha* and Microsoft's *ASIRRA* [31].

A Captcha can also be described as a reverse Turing test in the sense that a Captcha is administered by a machine and targeted at a human, in contrast to the original Turing test where roles are switched. When a Captcha challenge is generated and presented by a service provider (server) to a service requestor (client), the server subsequently uses the submitted response in order to distinguish a human client from a machine client and grants or denies access, accordingly. As the server has to solve a classification task, two types of errors can occur: (1) a "false-positive" (FP) occurs when the client is misclassified as a machine although it is a human; (2) a "false-negative" (FN) occurs when the client is misclassified as a human although it is a machine.

When a FN occurs, a machine was capable of solving the challenge correctly and erroneously gets access to a service. This violation of authorization has been discussed widely in the security-related literature [1, 98, 14, 85] and recent studies indicate that many state-of-the-art, text-based Captchas have been broken [7, 12, 18, 25, 37, 56, 65, 83]. Consequently, the need for new (non text-based) types of Captchas is manifest, as suggested by several researchers [7, 12, 20, 54, 56].

When a FP occurs, a user was not capable of solving the Captcha task correctly and needs to continue solving Captcha tasks until s/he succeeds. This procedure can be tedious and may finally exclude the user from the requested service when s/he gives up. In contrast to FNs, this issue is not related to security but to usability. In fact, the objective of achieving both, high levels of security and usability, is a critical challenge in the design of

[1] For the sake of readability, we do not capitalize the term "CAPTCHA" in this paper.

Captchas [7, 15, 18, 33, 37, 58, 59, 69].

Shortcomings of past research on Captchas include a) a limited diversity of designs of image-based Captchas, b) a low number of empirical studies on the usability of image-based Captchas, and c) a low number of empirical studies which compare the usability of text-based and image-based Captchas. As a result, there is a lack of empirical insights which, in turn, limits the practical design and deployment of Captchas for service providers. Low levels of usability of a Captcha negatively affect the use and final adoption of the related service when the users get frustrated solving the Captcha and finally move to a service of a different provider. This is consistent with the findings of Chang and Chen [19], who show that - mediated by customer satisfaction - the quality of an interface (with a Captcha being a part of it) significantly affects customer loyalty.

In this paper, we contribute to research by (1) suggesting two new Captchas based on face recognition, (2) analyzing their security, and (3) empirically evaluating these Captchas against two established text-based Captchas with regard to four usability measures (effectiveness, efficiency, learnability and satisfaction). With regard to contribution (1), we developed two novel Captchas, which we refer to as *Farett-Gender* and *Farett-Gender&Age*[2]. The Captcha *Farett-Gender* presents a set of pictures, with each picture showing either a man or a woman, and requires a user to select all pictures showing a man/woman. The Captcha *Farett-Gender&Age* follows the same procedure and in addition, the user has to distinguish young, middle-aged and old persons of a given gender. Concerning contribution (2), we perform a combinatorial analysis, which accounts for two realistic kinds of attacks: brute-force attacks and attacks based on artificial intelligence. With regard to contribution (3), we conducted a within-subjects study where participants had to solve instances of both face recognition Captchas and instances of two state-of-the-art Captchas: *reCAPTCHA* provided by *G*oogle and the open source project *K*Captcha [77]. To the best of our knowledge, we provide the first empirical usability study that compares image-based Captchas with text-based Captchas.

The remainder of this paper is structured as follows: In the next section, we frame our research and present a brief literature review. Then, we present two novel Captchas, analyze their security and describe our empirical usability study before we show its results. We discuss our findings and conclude with a summary and an outlook.

[2] Farett is an acronym for **Fa**ce recognition **re**verse **T**uring **t**est.

2. Framing the discussion

In this section, we provide a brief overview of the various types of Captchas, operationalize the usability criteria for our usability study, and provide a structured presentation of literature findings on security and usability of Captchas.

2.1. Types of Captchas

The most common types of Captchas are text-based, audio-based, image-based or video-based [7, 44, 74, 98].[3] The *noCaptcha* version of *reCaptcha*, which has been introduced recently by Google, implements a different approach: *noCaptcha* analyzes how a user interacts with a website and verifies that s/he is human by having her/him click on the "I'm not a robot" checkbox[4]. In doubt, *noCaptcha* additionally prompts the user with the text-based *reCaptcha* as a fallback. Although the *noCatpcha* checkbox allegedly implements a "high degree of sophistication", it is not a classic challenge-response type of Captcha. The lack of technical information on how *noCaptcha* works creates an impression of "security by obscurity", which gives rise to legitimate skepticism among security researchers, e.g., [24]. Figure B.1 shows examples for each of the four established Captcha types.

Text-based Text-based Captchas require users to recognize distorted text rendered in an image. The security of these Captchas is based on the assumption that optical character recognition (OCR) algorithms fail to solve this type of problem. Google's *reCAPTCHA*, for instance, uses scanned texts from old books or newspapers, and lets the user decipher those words that could not be recognized by OCR algorithms. To enhance security, the presented texts are randomly distorted (see Figure B.1a).

A different approach for text-based Captchas is adopted in *SemCAPTCHA* [53]. It takes the otherwise primitive task of recognizing and typing letters one step further and adds a semantic element to it. *SemCAPTCHA* challenges the user to recognize three distorted words and to sort out which one does not fit semantically to the other two (e.g., one word corresponds to a reptile while the others are related to mammals).

[3] Other types of Captchas, such as hybrid combinations and interaction-based Captchas, are rarely addressed in the literature.

[4] http://googleonlinesecurity.blogspot.co.uk/2014/12/are-you-robot-introducing-no-captcha.html

Audio-based Originally designed to overcome accessibility problems of visually impaired people, audio-based Captchas rely on sound or speech recognition by humans. The audio sequence is generally composed of a set of words to be identified, layered on top of background noise as distortion (see Figure B.1b).

Many of today's audio-based Captchas are not considered to be user-friendly [13, 10]. For example, in a study with 89 blind participants, only 43% were successful in solving ten popular audio-based Captchas Bigham and Cavender [10].

Image-based Users are asked to perform an image recognition task, for example to distinguish an object or idea from a picture, by recognizing facial features of a human face, or to identify the upright orientation of an image. Image-based Captchas rely on the so called "semantic gap" between humans and machines, according to which humans naturally outperform algorithms in extracting information from given images [82]. One example for an image-based Captcha is Microsoft's *ASIRRA* (see Figure B.1c), which prompts users to identify all cats out of a set of twelve photographs of both cats and dogs. Another example is *Confident CAPTCHA* [23], which asks users to click on all images that show a given type of symbols, such as cats, birds or aircrafts.

Image-based Captchas, such as *ASIRRA*, have the particular advantage that, unlike text-based Captchas, they can be used across linguistic boundaries.

Interaction-based Captchas are a subtype of image-based Captchas. These Captchas address the human ability to logically perceive complex images and to perform a task based on perception, such as clicking on the mountaintop or putting a randomly rotated image back to its upright position [6, 42].

Video-based Video-based Captchas rely on the assumption that only humans can extract complex information out of video-clips. Thus, these Captchas prompt users to either describe in some words what happens in a video (see Figure B.1d) or to type in the slogan from a video advertisement.

2.2. Usability and security

Having defined usability as "how intuitive and easy the Captcha is for all users", we operationalize the construct of usability for the purpose of measurement, as previously suggested in ISO 9241-11[5] and in [44, 66, 98][6]:

1. **Effectiveness** Effectiveness usually subsumes both, FPs and FNs of Captchas. In the context of usability, only FPs are relevant. We refer to FPs as errors and to FP rates as error rates in the following. Ahmad et al. [1] suggest that humans should be able to successfully solve the Captcha at a rate of 90% or above, i.e., to ensure adequate usability, error rates should not exceed 10%.

2. **Efficiency** The time for solving a Captcha (either correctly or incorrectly) represents an important usability parameter. According to Rui and Liu [75], 30 seconds are considered to be the approximate limit beyond which the efficiency of a Captcha and therefore its usability declines dramatically.

3. **Learnability** This criterion refers to how easy it is for users to accomplish basic tasks when they encounter a design for the first time [67]. The more intuitive a Captcha is, the easier and faster it can be learned.

4. **Satisfaction** The user's overall satisfaction is a subjective indicator of his/her acceptance of the Captcha procedure. The acceptance, in turn, affects the user's willingness to use web services that grant access by means of the respective Captcha.

The security analysis of Captchas can address different aspects. The following security requirements are proposed in the literature [16, 94, 100]:

5. **Media security** Media (images, audio sequences etc.) should be obfuscated before being presented to the user. For example, texts and audio sequences should be distorted. Otherwise, Captchas become easily vulnerable to attacks performed with text recognition tools, speech recognition tools, etc. [14].

   Media security can be compromised through machine-learning attacks. Secure Captchas should prevent machine-learning attacks or at least make them hard to perform [100].

[5] This ISO standard suggests that usability measures should cover effectiveness, efficiency and satisfaction.

[6] Nielsen [66] also mentions *memorability* as a usability parameter. However, measuring memorability requires a long-term experimental setup where participants are tracked over a longer period of time. As we implemented a short-term experimental setup, memorability could not be tested.

6. **Script security** Captcha systems should ensure that there are no easy ways to solve Captchas at the script level. Common examples for compromised security are those where the answer to the Captcha challenge is passed in plain text or where a solution to a Captcha challenge can be used multiple times (the Captcha is then vulnerable to replay attacks).

7. **Algorithmic/randomness security** Captcha challenges must be generated and evaluated automatically [91]. For this automated process, Captchas rely on random algorithms. The literature shows that the security of Captchas depends on a strong algorithmic randomness of the Captcha challenges used [14, 100].

8. **Security after wide-spread adoption** Captchas should be secure even after a significant number of websites adopt them. This requirement may be violated, for example, when a text-based question is asked, such as a mathematical question like "what is 1+1". A parser could easily be written that would allow bots to bypass this test.

2.3. Literature findings

In order to identify the state-of-the-art of Captcha-related security and usability research, we performed a literature review. Our literature search process is described in Appendix A. The results of our review are presented in Table B.1, which shows the contributions of the literature for each of the five Captcha types presented in Subsection 2.1. We analyze each paper with regard to which particular usability and security criterion is covered.

In total, we found 18 papers that address usability issues of Captchas and 38 papers that are dedicated to security aspects. As Table B.1 indicates, research has focused on four clusters: 1) Security of text-based Captchas, 2) Usability of text-based Captchas, 3) Security of image-based Captchas and 4) Usability of audio-based Captchas. The usability of image-based Captchas has heretofore been largely neglected.

3. Development of two face recognition Captchas: ***Farett-Gender*** and ***Farett-Gender&Age***

In this section, we propose two novel face recognition Captchas (*Farett-Gender* and *Farett-Gender&Age*), which are based on gender and age classification. Before presenting the Captchas, we discuss the state-of-the-art of gender and age classification in the literature on Captchas.

3.1. Gender and age classification for Captchas

Both gender and age classification are challenging and interesting problems in the field of artificial intelligence (AI) and have attracted much attention by researchers in recent years, particularly with regard to the rise of surveillance technology [55].

Gender-based classification As acknowledged widely in the literature, the accuracy of gender recognition algorithms strongly depends on the quality of images. To be precise, gender recognition of uncontrolled, real-world images, i.e., of non-frontal, poorly lit faces, has been considered an unsolved problem in AI research [4, 38, 79]. The security of *Farret-Gender* is based on this gap of accuracy between machines and humans in recognizing gender. To estimate how well machines would perform on *Farret-Gender*, we consider studies that evaluate the accuracy of gender recognition algorithms on real-world images. Unfortunately, most studies classify faces taken under controlled conditions: for example, Mäkinen and Raisamo [55] test methods for automated gender classification and achieve a gender recognition rate of 89% on average. However, it should be noticed that a face database was used with pictures taken under controlled lighting and positioning conditions. In another study, Shan [79] achieves a recognition rate of up to 94.81% with an improvement of current classification algorithms. The study is based on a picture database which was previously cleared of unfavorable images, such as not (near) frontal ones. According to Shan, this step was necessary because otherwise the classification algorithms would have performed much worse. Studies classifying more realistic images use images which are not publicly available [38, 78] or they do not transparently select a subset of publicly available images [79]. While many of these studies show how well algorithms can perform on high quality images taken in controlled environments, some studies also indicate under which circumstances the algorithms perform suboptimally. For example, the recognition rates drop to approximately (i) 60% for some ethnicities [38], (ii) 70% for a majority of the age groups [26], (iii) 70-80% for female subsets [30] and (iv) 75% for more realistic images [78]. There is further evidence for the difference between humans and machines in recognizing gender: those faces that were misclassified [26] or excluded [79] in the studies could easily be classified by humans. On the one hand, some authors report recognition rates in excess of 90% [79] and recent studies show that for some real-world images, similar success rates are possible [17, 86].

On the other hand, the studies and examples presented above indicate that current AI-based methods must be expected to perform worse if the images presented by *Farret-Gender* are not taken under controlled conditions or if they are distorted, for example.

To determine how accurately gender recognition tools classify the images used in our study, we selected the state-of-the-art gender recognition tool provided by *Project Oxford*[7] and classified a set of 500 (undistorted) images from *Labeled Faces in the Wild* [87, 48, 47], which includes all images used in our usability study. We also classified a set of 220 frontal faces from the *Color FERET Database* [72, 71], which contains high-quality images taken under controlled conditions. The gender recognition tool achieved a success rate of 87.6% on the images from *Labeled Faces in the Wild* and 96.8% on the images from the *Color FERET Database*. These results provide further evidence and show that machines are much less accurate in gender recognition when low-quality and uncontrolled images are used (compared to classifying high-quality and controlled images). This supports our assumption that, in the presence of uncontrolled images, humans capability of gender recognition is still superior to that of machines.

Age-based classification Geng and Smith-Miles [39] show that age classification can be performed by algorithms with a mean absolute error of 5.36 years. Human evaluators achieved a mean absolute error of 8.13 (only gray-scale face regions are shown) and 6.23 years (whole color images are shown) in the same test environment. However, the interpretation of these results needs to consider that a) these experiments, again, rely upon databases with face pictures that were mostly taken under controlled conditions, including the use of frontal images, and b) no information on the human evaluators is provided (e.g., how much time did they have?) and c) only 29 human evaluators participated. We assume that under realistic, uncontrolled conditions, algorithms perform worse than reported by Geng and Smith-Miles [39]. Horng et al. [45] suggest an age group classification method based on neural networks, which shows an average success rate of $q = 78.13\%$ for gray-scale facial images when three age groups ("young", "middle" and "old") are used (p. 190); further studies are reported by Fu et al. [36], who indicate that age classification of non-frontal images remains

[7] https://www.projectoxford.ai/demo/face#detection

an open problem.

We assume that the human brain is superior to computers in performing age and gender classification tasks in terms of accuracy. Thus, we propose two novel Captchas, *Farett-Gender* and *Farett-Gender&Age*, which are based on gender and age classification. We then analyze their security and usability.

3.2. Farett-Gender and Farett-Gender&Age

Figure B.2 shows the Captcha *Farett-Gender*[8], which is based on gender classification. It prompts the user with faces ($n = 20$), among which exactly $k = 5$ faces are corresponding to the same gender, which is selected arbitrarily. The user has two options: s/he can either click on five images and then press the submit button, or s/he can press the reload button to get a new set of pictures. The option to request a new challenge is widely used in today's Captchas.

Our second Captcha, *Farett-Gender&Age*, requires the user to sequentially perform two classification tasks, gender and age classification. It prompts the user with $n = 16$ images. In a first step, the user needs to click on three images of a given gender (see Figure B.3). The user's options are similar to those available for *Farett-Gender*. If task 1 has been completed correctly, the user needs to mark the youngest person of the three selected pictures (see Figure B.4), then s/he need to mark the oldest one. The order of ages varies and is determined randomly.

The number of pictures used in the *Farett* Captchas is based on the following rationale. First, we considered the implementation of similar, established Captchas. For example, ASIRRA uses a number of twelve different pictures. With this set of images, the authors of ASIRRA claim their Captcha to have a combinatorial breaking probability of 0.2% [31]. Considering an attack based on learning classification algorithms increases this probability to 10.3%, according to Golle [40]. Thus, the number of twelve images can be considered as a minimum with which a sufficient level of security can be ensured. In addition, we took into account the feedback of a small pre-test group of six participants, who were asked to evaluate *Farett* Captchas which showed different numbers of images regarding usability. Based on their feedback, the number of pictures should not exceed

[8] The experiments were conducted in German. All screenshots in this paper show translated versions.

25. We decided to use for both *Farett* Captchas a single number of images for the following reason: As we apply a within-subject design in our study, using several values for the number n of images, or even the entire range, would probably lead to a high drop-out rate because of the efforts required by the participants. We decided to select $n = 20$ and $n = 16$ for *Farett-Gender* and the more complex *Farett-Gender&Age*, respectively, with these values being a trade-off between a low drop-out rate and an acceptable security level.

4. Security of *Farett-Gender* and *Farett-Gender&Age*

Regardless of the usability of our Captcha procedures, they have to be sufficiently secure. We evaluate how secure our proposed Captchas are with regard to those two types of attacks which are commonly considered in the literature [7; 99, p.158]: brute-force attacks and attacks based on artificial intelligence (AI).

In the following analysis, we consider an attacker who solves a single instance of a particular Captcha C with probability θ_C. Trying to solve any of i given Captcha instances, an attacker succeeds with probability

$$p = 1 - (1 - \theta_C)^i \,. \tag{1}$$

Captchas which can be solved by a machine - i.e., by a brute-force or AI-based attack - with a probability in excess of 1% are commonly considered insecure in the literature [14, 20, 37].

Let us now consider the security of Google's *reCAPTCHA* - both, as an example, and as a benchmark for our Captchas. When a Captcha challenge presents a text of length n with each single character drawn from a set of m characters, the probability of success of a brute-force attack amounts to

$$\theta_{reCAPTCHA} = \frac{1}{m^n}. \tag{2}$$

For example, if a text contains seven characters, with each character being either a digit, a lower case letter or an upper case letter, then the attacker's probability of guessing correctly equals $\frac{1}{62^7} \approx 2.84 \cdot 10^{-13}$. Even without considering the challenge of guessing the number of characters presented by the Captcha, this is well below the threshold of 1%.

With regard to AI-based attacks, a wide range of success rates has been reported for different versions of *reCAPTCHA*: 4.9% [20], 11.6% [5], 24.7% [29], 31% [46], 40% [25, 83], and even 55% [83].

4.1. Farett-Gender

In a brute-force attack on *Farret-Gender*, an attacker has to guess k images out of a set of n images, from a combinatorial perspective. Thus, s/he needs to guess one solution out of $\binom{n}{k}$ options, resulting in an attacker's success probability (FN rate) of

$$\theta_{Farret-Gender} = \frac{1}{\binom{n}{k}} = \frac{k!(n-k)!}{n!}. \qquad (3)$$

This amounts to approximately $0.00645\% = 6.45 \cdot 10^{-5}$ for $n = 20$ and $k = 5$, which can be considered secure.

When an attacker uses an AI-based gender determination algorithm to break a *Farret-Gender* instance, s/he does not need to determine the gender of each of the $n = 20$ persons correctly in order so succeed. For example, when an attacker applies the AI-based algorithm to each of the 20 images and yields $k = 6$ images which are classified as those of a given gender, five of these images may be correctly classified and one misclassified. In this case, the attacker can guess which of the six images was misclassified, resulting in a chance of about 16.67%. In our analysis, we assume that sophisticated AI-based algorithms are robust with regard to gender determination, i.e., the probability p with which the algorithm correctly determines the gender of a person shown on an image may vary slightly but does not substantially depend on the particular image. As a consequence, sorting probabilities computed for the n images in descending order and selecting the first k images is not promising. Assuming a constant probability p for automatically determining the gender of a person shown on an image, allows to compute the overall probability $\theta'_{\text{Farett-Gender}}$ with which an attacker succeeds in breaking a *Farret-Gender* instance using an AI-based gender determination algorithm. Addressing the general case (correctly select k out of n images), we yield

$$\begin{aligned}\theta'_{Farett-Gender} \quad = \quad & \frac{1}{\binom{n}{k}} \cdot \left[\sum_{i=0}^{k-1} (1-p)^{k-i} \cdot p^{n-k+i} \cdot \binom{n}{i} + \right. \\ & \left. \sum_{j=0}^{n-k} (1-p)^{n-k-j} \cdot p^{k+j} \cdot \binom{n}{j} \right]. \qquad (4)\end{aligned}$$

It should be noticed that, for $p = 1$, $\theta'_{Farett-Gender}$ equals 1. A proof of equation (4) is provided in Appendix B. For different accuracy levels (p) of

AI-based gender detection algorithms, Tables B.2-B.4 show how the parameters n and k need to be adjusted in order to keep the overall probability of a successful attack on *Farret-Gender* under a given threshold. Taking into account the state-of-the art of gender-recognition algorithms (see Subsection 3.1), we perform computations for $p = 80\%, 85\%$ and 90%.

The tables show that, for each value of p, the success probabilities for the cases $(n = 20, k = 9)$ and $(n = 20, k = 11)$ are identical as a) in the latter case 11 images of a given gender, or 9 images of the opposite gender, have to be identified and b) the probability of correctly identifying the gender of a person does not depend on the particular gender.

In our experimental setting $(n = 20, k = 5)$, for example, the success probability amounts to 2.54% for $p = 80\%$; in order to decrease the success probability to a value lower than an acceptable 1% (see discussion above), the numbers of images shown and images to select (n, k) should be slightly increased to $(26, 5), (24, 7), (23, 9)$, or $(23, 11)$, with n being at an acceptable level of about 25. With $p = 85\%$ the numbers of n and k would need to be increased to $(37, 5), (33, 7), (32, 9)$, or $(31, 11)$; with $p = 90\%$ the numbers of n and k would need to be increased to $(57, 5), (52, 7), (49, 9)$, or $(48, 11)$. In both cases, the number n of images needed to keep the attacker's success probability below 1% exceeds 25.

To sum up, our analysis reveals that *Farett-Gender* is secure with regard to brute force attacks. Concerning AI-based attacks, algorithms showing a gender-recognition success rate of above 80% would threaten the security of *Farett-Gender* as the number of images presented to a user should presumably not exceed 25 based on the feedback of a small pre-test group of six participants (see Subsection 3.2). Due to the very small size of this group, the maximum number of pictures that will be accepted by users would need to be substantiated and distinguished according to user characteristics, such as Internet and computer experience, age, cultural background etc., in future research. Unless the accepted number of pictures cannot be increased to a level of more than 30 pictures, e.g., by other ways in which images are presented to the user, the proposed Farett Captchas cannot be both secure and usable. Yet, beyond the option to increase the number of pictures, there are several countermeasures for maintaining and improving the robustness against AI-based attacks, such as image-distortion and deidentification [28, 9]. However, new empirical studies beyond those in this work would be required in order to assess the usability of such *Farett-Gender* instances.

4.2. Farett-Gender&Age

The Captcha *Farett-Gender&Age* can be subject to a more diverse range of attacks and therefore requires a more elaborate security analysis. We assume in the following that the number k of images to select and finally to sort according to age equals 3. We believe that users are not willing to sort more than three images as sorting is a more challenging task than separately classifying images (as required in the gender task). However, our analysis can be straightforwardly extended to arbitrary values of k.

From a conceptual perspective, either of the tasks of classifying age and gender can be conducted through guessing (brute force) or through using AI methods, resulting in four combinations:

1. (Gender: brute-force, age: brute-force) Drawing on our security analysis of *Farett-Gender*, the probability of guessing a *Farett-Gender&Age* instance correctly amounts to

   $$\theta_{Farett-Gender\&Age} = \frac{3!(n-3)!}{n!} \cdot \frac{1}{3!} = \frac{(n-3)!}{n!}. \tag{5}$$

   Having $n = 16$ images in our experimental setting, an attacker's success probability is approximately 0.03%, which is acceptably low.

2. (Gender: brute-force, age: AI-based)

   As in the case of gender classification, AI algorithms have been proposed and tested for age classification (see Subsection 3.1). Let q denote the probability with which an AI algorithm correctly assigns an image to an age group (e.g., "young" or "old"). The result of ordering three images according to their age group can lead to 2^3 cases as each of the 3 images is either classified correctly or not. One case is characterized by correctly classifying all three images; this case occurs with probability q^3.

   In $\binom{3}{1} = 3$ cases, exactly one image is misclassified. The probability of each of these cases equals $q^2 \cdot (1-q)$. Now the attacker faces a case in which two images are assigned to the same age group. S/he has to guess and we assume that s/he guesses correctly with a probability of 50%. Overall, the success probability of the cases amounts to $\frac{3}{2} \cdot q^2 \cdot (1-q)$.

   In $\binom{3}{2} = 3$ cases, exactly two images are misclassified. The probability of each of these cases equals $q \cdot (1-q)^2$. Without loss of generality, we assume that the attacker misclassifies the images that show the young and the old person; there are 4 possibilities of this misclassification (we assume they are equally distributed) as each of the 2 images can be misclassified in 2 ways. Only when both images are misclassified as “middle”, the attacker can succeed. In this situation, the attacker has to guess the correct assignment of images to age groups out of $3! = 6$ possible assignments. Overall, the success probability of the cases amounts to $\frac{1}{8} \cdot q \cdot (1-q)^2$.

Finally, in one case all three images are misclassified with probability $(1-q)^3$. In this case, the attacker fails regardless of the age groups assigned to the three images.

To sum up, using an AI-based age classification of three images an attacker succeeds with probability

$$q^3 + \frac{3}{2} \cdot q^2 \cdot (1-q) + \frac{1}{8} \cdot q \cdot (1-q)^2. \tag{6}$$

Overall, the probability of guessing a *Farett-Gender&Age* instance correctly amounts to

$$\theta'_{Farett-Gender\&Age} = \frac{3!(n-3)!}{n!} \cdot \left(q^3 + \frac{3}{2} \cdot q^2 \cdot (1-q) + \frac{1}{8} \cdot q \cdot (1-q)^2 \right). \tag{7}$$

Having $n = 16$ images and assuming a very high accuracy of AI-methods ($q = 95\%$), an attacker's success probability is approximately 0.17%, which is acceptably low.

3. (Gender: AI-based, age: brute-force) Assuming that gender classification can be performed with accuracy p, an attacker can break a single *Farett-Gender&Age* instance with probability

$$\begin{aligned} \theta''_{Farett-Gender\&Age} &= \theta'_{Farett-Gender} \cdot \frac{1}{3!} \\ &= \frac{(n-3)!}{n!} \cdot \left[\sum_{i=0}^{2} (1-p)^{3-i} \cdot p^{n-3+i} \cdot \binom{n}{i} + \right. \\ &\qquad \left. \sum_{j=0}^{n-3} (1-p)^{n-3-j} \cdot p^{3+j} \cdot \binom{n}{j} \right]. \end{aligned} \tag{8}$$

Table B.5 shows the probabilities with which an attacker succeeds.

The values in Table B.5 show that in our experimental setting ($n = 16, k = 3$), the success probability amounts to 1.26% for $p = 80\%$; in order to decrease the success probability to a value lower than an acceptable 1%, the number n of images should be increased to 18. For more accurate AI-based gender detection algorithms ($p = 85\%$ and $p = 90\%$), n would need to be increased to 25 and 40, respectively.

4. (Gender: AI-based, age: AI-based) When an attackers draws on AI-based methods for both gender classification and age classification, the success rate per *Farett-Gender&Age* instance amounts to

$$\begin{aligned}\theta'''_{Farett-Gender\&Age} &= \theta'_{Farett-Gender} \cdot \left(q^3 + \frac{3}{2} \cdot q^2 \cdot (1-q) + \frac{1}{8} \cdot q \cdot (1-q)^2\right) \\ &= \frac{1}{\binom{n}{3}} \cdot \left[\sum_{i=0}^{2} (1-p)^{3-i} \cdot p^{n-3+i} \cdot \binom{n}{i} + \right. \\ &\qquad \left. \sum_{j=0}^{n-3} (1-p)^{n-3-j} \cdot p^{3+j} \cdot \binom{n}{j}\right] \\ &\qquad \cdot \left(q^3 + \frac{3}{2} \cdot q^2 \cdot (1-q) + \frac{1}{8} \cdot q \cdot (1-q)^2\right). \qquad (9)\end{aligned}$$

In our case ($n = 16$), the success rate of an attacker amounts to approximately $5.35\%, 11.33\%$ and 23.75% for $p = q = 80\%$, $p = q = 85\%$ and $p = q = 90\%$, respectively. Table B.6 shows the probabilities with which an attacker succeeds.

In order to decrease the success probability to a value lower than an acceptable 1%, the number n of images shown needs to be increased to 31, 45 and 73 for $p = q = 80\%$, $p = q = 85\%$ and $p = q = 90\%$, respectively. When users are not willing to classify more than 25 images, we conclude that for $p, q \geq 80\%$, *Farett-Gender&Age* is insecure against AI-based attacks on both gender and age classification.

To sum up, our analysis of *Farett-Gender&Age* reveals that it is (1) secure against brute force attacks on both gender and age recognition, (2) secure against brute force attacks on gender recognition and "realistic" AI-based attacks on age recognition (even when the age-recognition rate amounts to a rate as high as 95%), secure against against AI-based attacks on gender recognition and brute force attacks on age recognition if the gender recognition rate p does not exceed 85% and the number of images shown is increased to 25, and secure against AI-based attacks on both gender and age classification unless age- and gender-recognition rates exceed 80% and users are not willing to classify more than 25 images.

It should be noticed that regarding age classification with AI-based algorithms, beyond the option to separately assign each of the presented images to one age class and then to select the youngest or oldest person, an attacker can also look for AI-based algorithms which rank a given set of faces according to the age and then select the oldest or youngest person based on comparisons. However, we are not aware of any study that analyzes how well this task can be performed with AI-based algorithms.

5. Usability study of the proposed Captchas

The main purpose of our study is the empirical evaluation of the proposed *Farett* Captchas against each other and against state-of-the-art Captchas in terms of usability. In order to keep the efforts for the participants of our study manageable and the drop-out rate low (this issue becomes particularly important when using a within-subject design, which we use for the advantages explained below), we decided to limit the number of Captchas to four. In addition to the two proposed Captchas, we use the two Captchas *reCAPTCHA* (see Figure B.1a) and *KCAPTCHA* (see Figure B.5) as benchmarks for two reasons:

1. Both Captchas are well-known, text-based reverse Turing tests. *reCAPTCHA* currently the most widely used Captcha, serving about 100 million Captchas instances each day [41].

2. For both Captchas, a public interface is provided, which allows for an efficient integration into our experimental environment.

We would have preferred *noCaptcha*, which was introduced recently by Google, as a benchmark for usability. However, this was not possible for two reasons: (1) Our pretests showed that it is impossible to reliably have participants solve more than two instances of *noCaptcha*. The reason is that users are prompted with the original, text-based *reCaptcha* as a fallback when they solve multiple instances of *noCaptcha*. This prevents us from gathering enough data for determining reliable statistics on usability, the time required for solving the Captcha and its error-rates. (2) Furthermore, as *noCaptcha* is not a Turing test in a sense that it implements a challenge-response procedure, it is impossible to isolate the time needed to solve the Captcha, i.e., to measure its efficiency.

To compare the four Captchas, we follow the recommendation of Nielsen [66, pp. 178] and implement a within-subject design in which every participant has to solve all four Captchas. The advantage of a within subject design is that it yields a higher statistical power than a between-subject design with the same number of participants[9]. In addition, the error variances associated with individual differences are reduced. We assume that carryover effects[10] - i.e., practice and fatigue, which are common shortcomings of within-subject designs - do not seriously affect our results. In general, the practice effect might mean that the participants are more confident and accomplished after the first test because they gained experience. In our case, practice is unlikely to be an issue, as all of the four Captchas pose straightforward tasks to the participants and, according to our pretest, we expected the participants of our final study to be well familiar with PCs, Internet usage and Captchas in general. With regard to fatigue[11], we argue that this is not a severe issue because in our pilot study, most of the participants took just a few minutes to complete all tests [12]. To sum up, our study was neither intellectually challenging nor time-consuming for participants so that carryover effects, if any occur, are probably minor. To control carryover effects, we change the order in which the four Captcha procedures are presented to the participants (counterbalancing, cf. Subsection 5.3).

As participants needed only a web browser, it was possible to conduct the study as a pure online test. Although online tests do not provide the same controlled environment as laboratory experiments, their advantages are an easier acquisition of participants, no need for laboratory space, and a more natural environment, which ensures a higher validity of results.

We conducted a pilot study with six participants and we interviewed them in order to test the design of the web-based study and the design of

[9] We would need to get twice (four times) as many participants as in the within-subject design if we tested the established Captchas and the newly proposed Captchas in two groups (each of the four Captchas in a different group).

[10] A carryover effect occurs when preceding tests adversely influence succeeding tests; i.e., the order of tests has an impact on results.

[11] Participants may become exhausted, bored or simply disinterested after taking part in multiple tests.

[12] In our empirical study, the mean time was 4:17 min, with a standard deviation of 1:15 min.

our *Farett* Captchas, including the number of images that are presented.

In the remainder of this section, we describe details of our study with regard to the participants, the usability measures, the image data, and the protocol[13].

5.1. Participants

We recruited 151 participants using the following communication channels:

- the mailing list of the WKWI, which is the German-speaking information systems community,
- Facebook groups of information system students,
- the web sites of the authors' research group and of our Faculty,
- lists of email addresses of information system students at the University of Regensburg, and
- colleagues of the authors.

We removed datasets from participants who either did not complete the test (26 participants) or took extremely long (7 participants) so that their results were assumed to be invalid. The latter were treated as outliers and removed when the overall time they needed to complete the test exceeded 209 seconds (μ+3×σ) [13]. This resulted in a sample of $N = 118$ participants (96 men and 22 women), aged between 17 and 60.

Participation was voluntary and no compensation was provided to attract only those participants who were interested in the study itself rather than in getting benefits. In the survey, participants were not only asked for demographic information but also for their experience with computers. Regarding the time, participants spend on the Internet (per day), we found an almost uniform (discrete) distribution up to seven hours, with 13 participants (11%) spending more than seven hours online. The perceived computer related knowledge of many participants was high (91% rated their knowledge between 7 and 10). Apparently, most of the participants are young and male adults with good computer skills. Table B.7 shows characteristics of the participants.

[13] Our institution does not require Institutional Review Board approval for human-based studies.

[10] x-y means: x hour(s) $\leq$ Internet usage $<$ y hours.

[11] One participant did not answer the question.

5.2. Image data

For both Captcha methods, we used the database *Labeled Faces in the Wild* [87, 48, 47], which is constantly growing. When we conducted the study, this database contained a total of 13,233 color portraits of various people of all ages. All images in the database were randomly collected from the Internet using the *Viola-Jones* face detection algorithm [88]. The security of *Farett* benefits from this random selection because shooting conditions and face rotation vary greatly.

5.3. Protocol

The study consists of three phases and it was designed as a web-based online test; we hosted all web pages and server applications on our own servers. In the first phase, we welcomed and introduced the participant on a starting page, which briefly explains the purpose of the study, provides the estimated time required for finishing the test (five minutes, based on our pretest results), informs about the tasks s/he has to solve (solve five instances of each of four different Captchas), and provides links to a short introduction to Captchas[12] and to the websites of *reCAPTCHA*[13] and *KCAPTCHA*[14]. However, following these links is not mandatory to be able to do understand the study and the required tasks. The main purpose is to provide background information to those who are interested.

In the second phase, the participant has to solve five instances of each of the Captcha types *reCAPTCHA*, *KCAPTCHA*, *Farett-Gender*, and *Farett-Gender&Age*, with different orders being used (counterbalancing). We apply the *4x4 Latin Square* design, as suggested by Mitchell and Jolley [60, p. 542], according to which participants are assigned to one of the four groups shown in Table B.8. The assignment of participants to groups follows a cycle: participant 1 is assigned to group 1, participant 2 to group 2 etc., and participant 5 to group 1 etc. The instances of all Captcha types are generated randomly so that participants get different Captcha instances. Using different instances for each participant improves robustness of results against any bias introduced through a (too) small set of Captcha instances.

A participant first has to solve five instances of the Captcha type of his/her group, then s/he is asked to evaluate the Captcha. Based on the suggestions of Penninger et al. [70] and Nielsen [66], s/he has to respond to

[12] http://de.wikipedia.org/wiki/CAPTCHA

[13] http://www.google.com/recaptcha

[14] http://www.captcha.ru/en/kcaptcha

each of the following questions on a scale from 1 (very bad / hard) to 10 (very good / easy).[15]

1. How easy was it for you to solve this Captcha?
2. As how intuitive did this Captcha appear to you?
3. What is your overall impression about this Captcha?

The participant then has to solve five instances of the second, third and fourth Captcha type of his/her group and answer the usability questions presented above. For each Captcha instance, the participant has the option to either solve the instance or to load a new instance; this option is usually available for state-of-the-art Captchas.

We now describe how each of the four Captcha types are presented to the participants. For the generation of *reCAPTCHA* instances, we used the web interface provided by Google, which randomly generates Captchas. Figure B.6 shows a screenshot of a *reCAPTCHA* instance as it is presented to the participant.

For *KCAPTCHA*, we used the free Captcha generator provided on the project website.[16] Figure B.7 shows a screenshot of a *KCAPTCHA* instance as it is presented to the participant.

In contrast to the previously tested Captchas, which are probably known to many participants, *Farett-Gender* is new to the participant and needs to be explained. Thus, we provided a text description and the opportunity to watch a short demonstration video (35 seconds) that shows how *Farett-Gender* works. We then asked the participant to solve instances of *Farett-Gender* (see Figure B.2).

The generation of *Farett-Gender* instances proceeds as follows: From the database *Labeled Faces in the Wild* (cf. Subsection 5.2), we randomly selected 89 male and 89 female pictures and manually classified them as either male or female. We also manually classified all 178 pictures as either “young”, “middle”, “old” or “no class”, with the latter being used for those

[15] The experiments were conducted in German. All questions reported in this paper are translated versions. We decided to not use the System Usability Scale (SUS) suggested in [11] with its ten questions for two reasons: First, using ten questions for four Captchas would each user require to answer 40 questions, in addition to solving several Captcha instances. We believe that these efforts would have increased the drop-out rate of users participating in our experiment substantially. Second, not all questions of SUS are applicable to our experiment. For example, the question “I found the various functions in this system were well integrated.” is meaningless in our experimental setting.

[16] See http://www.captcha.ru/en/kcaptcha/

images where age classification might lead to ambiguous results. Overall, we classified 18, 21 and 25 images showing young, middle and old male persons, respectively, and we yielded 17, 28 and 20 images showing young, middle and old female persons, respectively. Based on these pre-classified images, *Farett-Gender* instances are generated in four phases (all of the selections are conducted at random): (I) a gender is selected, (II) from the set of 89 images of the selected gender, five images are selected, (III) the remaining 15 images are selected from the opposite gender and (IV) the 20 images are arranged in a rectangle.

For *Farett-Gender&Age*, we provided a textual explanation as well as the option to watch a short demonstration video (24 seconds) that shows how *Farett-Gender&Age* works. We then asked the participant to solve instances of *Farett-Gender&Age* (see Figures B.3 and B.4). The generation of *Farett-Gender* instances is similar to the procedure described above (all of the selections are conducted at random): (I) a gender is selected, (II) for the selected gender and each of the age groups "young", "middle" and "old", one image is selected, (III) the remaining twelve images are selected from the opposite gender, regardless of age and (IV) the 15 images are arranged in a rectangle.

In the third and final phase, the participant completes a questionnaire containing demographic items and measures pertaining to Internet usage and PC knowledge as suggested by Fidas et al. [34].

5.4. Measures

As suggested in Subsection 2.2, we use the following metrics to evaluate both, the objective performance and the perceived quality of the Captchas:

Effectiveness To measure effectiveness, we track the error rate of participants for each Captcha; reloaded (and thus unsolved) Captchas do not affect error rates. The error rates are defined as follows:

$$error\ rate = \frac{\#errors}{\#solved}$$

where #solved is the number of instances presented to the participant (in our case, five instances).

Efficiency To measure efficiency, we track the time a participant needs to solve a Captcha instance; more precisely, we measure the time period between the points of time when the instance is presented and the participant clicks on the "Submit" button. Thus, the time for reloading Captcha instances is not included.

Learnability In Section 2.2, we define learnability as how easy it is for users to accomplish a task when they encounter a design for the first time. Since it is difficult to measure learnability directly, we ask participants for their subjective impression (on a scale of 1 (poor) to 10 (excellent)); they had to answer two questions: *How easy was it for you to solve this Captcha?, As how intuitive did this Captcha appear to you?* We then calculate the mean of both values.

Satisfaction We measure satisfaction by asking the participants to express their overall impression about a certain Captcha on a scale from 1 (poor) to 10 (excellent); we use a third question as already mentioned in the protocol description: *What is your overall impression about this Captcha?* This measure can be considered as an aggregate of perceived effectiveness, perceived efficiency, and perceived learnability.

6. Usability results

We evaluate the tested Captcha procedures (*Farett-Gender*, *Farett-Age&Gender*, *reCAPTCHA* and *KCAPTCHA*) with regard to the four usability criteria (effectiveness, efficiency, learnability and satisfaction). For each of these criteria, we have one sample (of size $N = 118$) per Captcha procedure. The Shapiro-Wilk test showed that the samples (over all four criteria) are not normally distributed (at the .01-level). As a consequence, we applied the nonparametric Wilcoxon matched-pairs signed-ranks test [80, pp. 609] to compare the criteria. We use the one-tailed test in those cases where the distribution of difference scores in the populations represented by the two samples is not symmetric about the median of the population of different scores. Otherwise, two-tailed tests were conducted. Table B.9 provides an overview of the statistical results. To avoid confusion, we show only results of one-paired tests in Table B.9 - the (few) results of two-paired tests are reported in the text.

6.1. Effectiveness

We assess effectiveness in terms of error rates (cf. Figure B.8). As each participant had to solve exactly five instances of each Captcha, error rates get values of the set $\{0, 0.2, 0.4, 0.6, 0.8, 1\}$. Wilcoxon matched-pairs signed-ranks tests indicate that *KCAPTCHA* has higher error rates than the other Captchas. *Farett-Gender&Age* shows much higher error rates than *reCAPTCHA* and *Farett-Gender*. Only two-tailed Wilcoxon matched-pairs signed-ranks tests could be conducted; they indicate differences between the error rates of *Farett-Gender&Age* and *Farett-Gender* ($V = 549.5, p < .001$) and between *Farett-Gender&Age* and *reCAPTCHA* ($V = 936.5, p < .001$).

6.2. Efficiency

We assess efficiency in terms of task completion times (cf. Figure B.9). Although Figure B.9 suggests that the number of outliers is high, the percentage of outliers does not exceed 5% (for every Captcha, there are 590 observations: 118 participants·5 instances/participant). Furthermore, there is no systematic pattern which distinguishes the outliers of the four Captchas. Wilcoxon matched-pairs signed-ranks tests indicate that *KCAPTCHA* instances were solved quicker than *reCAPTCHA* ($V = 5185, p < .001$) instances. These were solved quicker than *Farett-Gender* ($V = 1044, p < .001$) instances, which, in turn, were solved quicker than *Farett-Gender&Age* ($V = 2551.5, p = .005$) instances. We found the same relationships of task completion times when we consider only times of successfully solved Captcha instances. With regard to task completion times of unsuccessfully solved Captcha instances, relationships are less clear: Wilcoxon matched-pairs signed-ranks tests indicate that participants require less time when they (unsuccessfully) try to solve *KCAPTCHA* instances compared to the time they require to (unsuccessfully) solve instances of any other Captcha procedure. The superiority of *reCAPTCHA* over *Farett-Gender* is at the .04 level of significance but the sample size ($N = 7$) is low.

6.3. Learnability

Regarding learnability, which was calculated as the average of perceived ease and perceived intuition, results are shown in Figure B.10. Wilcoxon matched-pairs signed-ranks tests indicate that *Farett-Gender* shows the highest learnability[17], and *KCAPTCHA* shows the lowest[18].

6.4. Satisfaction

Regarding satisfaction, results are shown in Figure B.11. As in the case of learnability, Wilcoxon matched-pairs signed-ranks tests indicate that *Farett-Gender* performs best. Specifically, *Farett-Gender* dominates *reCAPTCHA*

[17] Specifically, *Farett-Gender* dominates *reCAPTCHA* ($V = 2052, p = .007$), *KCAPTCHA* ($V = 950, p < .001$) and *Farett-Gender&Age* ($V = 3724, p < .001$) with regard to learnability.

[18] *KCAPTCHA* is dominated by *reCAPTCHA* ($V = 579, p = .004$), *Farett-Gender* ($V = 56, p < .001$)) and *Farett-Gender&Age* ($V = 224, p < .001$) with regard to learnability.

($V = 1719, p = .003$) and *KCAPTCHA* ($V = 797.5, p < .001$) with regard to satisfaction.

7. Discussion

7.1. Security

The results of our security analysis show that *Farett-Gender* and *Farett-Gender&Age* are secure against brute-force attacks. However, *Farett-Gender* is secure against AI-based attacks on gender recognition unless success rates exceed 80% and users are not willing to classify more then 25 images. While *Farett-Gender&Age* is secure against attacks where AI-based algorithms are applied to either age or gender recognition (as long as the success rate of a gender-recognizing algorithm does not exceed 85%), it is secure against AI-based attacks on both gender and age recognition unless age- and gender-recognition rates exceed 80% and users are not willing to classify more than 25 images.

In order to achieve security against AI-based attacks on *Farett-Gender* and *Farett-Gender&Age*, one option would be to increase the number of images that users have to classify. Another option would be to use images of lower quality (e.g., distorted images). As mentioned in Subsection 4.1, in both cases new usability studies would be necessary as the conditions under which users have to classify images change in terms of image quality and/or numbers.

Besides these essential types of attacks, a few common security issues should be anticipated when face recognition Captchas get adopted widely. (i) When face-recognition Captchas are employed at large scale, applying distortion techniques becomes necessary to counter attacks based on manually pre-classifying images and using hashing algorithms to match the images presented by the Captcha with pre-classified images. In the literature, such distortion techniques have already been applied to the image database, *Labeled Faces in the Wild*, which was used in our implementation of *Farett-Gender* and *Farett-Gender&Age* [9]. (ii) When the web server gets a high number of submissions during an unusually short time period, there are well established mechanisms to counter obvious attacks. First, the number of Captcha submissions per IP address and time period could be limited. Second, the submission of many Captcha solutions can be made more costly by requiring the client to perform time-consuming calculations known as hashcash [7]. Third, the web server could temporarily switch to a different Captcha procedure that is more robust against guessing even if it is less user-friendly.

7.2. Usability

Our usability study compares the proposed face recognition Captchas with text-based Captchas in terms of their objective quality (effectiveness and efficiency) and in terms of their subjective quality (learnability and satisfaction).

Considering effectiveness in terms of average error rates, we found that *Farett-Gender*, *reCAPTCHA* and *Farett-Gender&Age* dominate *KCaptcha*. Our results also indicate that there is no difference between the error rates of our face recognition Captchas and *reCAPTCHA*. However, we would not consider this finding conclusive, especially since other studies have reported substantially higher error rates for *reCAPTCHA*. For example, Bursztein et al. [13] measured an error rate of 25% in a large-scale study based on Amazon Mechanical Turk, which suggests that further investigations might confirm stronger, or even significant differences between the error rates of *Farett-Gender* and *reCAPTCHA*.

The analysis of efficiency shows that participants needed more time for solving our Captchas than for solving *reCAPTCHA* and *KCAPCTHA*. There are two possible reasons why our Captchas are less efficient. First, high task completion times can be based on the time-consuming challenge for humans to determine gender and age of persons shown in a picture. Second, they can also be due to the specific design and required user interactions in our experiment. In this case, alternative designs may lead to an improved efficiency.

With regard to learnability, *Farett-Gender* performs best, showing that *Farett-Gender* is easy to use and intuitive. Interestingly, *Farett-Gender* performs significantly better than *reCAPTCHA*, although the latter should be known to most participants in contrast to *Farett-Gender*. *Farett-Gender&Age* shows similar values as *reCAPTCHA*. Apparently, *Farett-Gender&Age* is regarded less learnable than *Farett-Gender*, probably because of the additional age classification task.

The results of participants' satisfaction with the four Captchas show that the proposed Captchas are superior in this regard to the established ones. With an average of 7.53 in satisfaction, *Farret-Gender* was rated significantly better than *reCAPTCHA* (6.66) and *KCaptcha* (4.80). This result is particularly relevant when customers' satisfaction with the interface of a web service affects the customers' satisfaction with the web service itself. Interestingly, lower efficiencies of *Farett-Gender* and *Farett-Gender&Age* do not lead to an overall lower satisfaction with these Captcha procedures.

In summary, our usability study indicates that, while the results of objective performance measures are mixed, our Captchas are promising with regard to measures of perceived quality. They comprise tasks that are easy, intuitive and that do not require any extraordinary knowledge. This makes our Captchas applicable in a broad range of contexts. Face recognition Captchas do not require a specific language or set of symbols, which would constrain them to particular regions or speakers of particular languages - a considerable advantage over today's widely deployed text-based Captchas. For example, Google's *reCAPTCHA* uses Latin letters, which are hard to perceive for users in, for example, China, Japan, Russia and Arabian countries. The fact that our Captchas do not require a keyboard makes them suitable for mobile devices - if bandwidth is not an issue.

There are two settings, in which face recognition Captchas should not be applied. In certain cultural contexts, it might be considered inappropriate to show uncovered faces and both service providers and users might thus be unwilling to use face recognition Captchas. Finally, alternative Captchas should be provided for visually impaired users, who will find our Captchas difficult to solve.

8. Summary and outlook

In this study, we proposed two novel face recognition Captchas, *Farett-Gender* and *Farett-Gender&Age*. We showed analytically that the proposed Captchas are secure with regard to brute-force and AI-based attacks unless age and gender recognition rates are higher than 80% and users are not willing to classify more than 25 images. Similar to the security of Captchas in general, which is threatened by advances of AI-methods, the robustness of our face recognition Captchas is affected by advances of face recognition algorithms. Face recognition Captchas can be strengthened by increasing the number of images, applying image distortion techniques and/or selecting images which are difficult to classify.

We further evaluated the suggested face recognition procedures and two text-based Captchas (*reCaptcha* and *KCaptcha*) in an empirical usability study. To the best of our knowledge, this article describes the first empirical usability study that compares image-based Captchas with text-based Captchas. Our findings indicate that *Farett-Gender* is similar to the other Captcha procedures concerning effectiveness, but it is less efficient; it is superior with regard to learnability and satisfaction. The other proposed Captcha *Farett-Gender&Age* performs worse than *reCAPTCHA* with regard to effectiveness and efficiency and similarly well with regard to learnability and satisfaction. To sum up, the proposed image-based *Farett-Gender* performs better than the currently widely deployed *reCAPTCHA*. The other face recognition Captcha, *Farett-Gender&Age*, turned out to be less usable which could be attributed to the fact that it is more complex.

As usability is an important indicator for the acceptance and deployment of Captchas, we argue that our study should be used to intensify the design and evaluation of Captcha procedures that are based on face recognition. With regard to the application of the proposed Captchas in protecting real resources on the web, a few implementation issues remain to be addressed. Although the image data used to generate our Captchas was sufficient for our proof of concept study, deploying the Captchas in a large-scale commercial context would require a bigger set of images. In particular, these images should not be publicly available to avoid attacks based on manually classifying the whole set of images. In addition, the set of images in the database must be high as, otherwise, an attacker can collect as many images as possible, label those by hand and automatically use the labeled images in subsequent Captcha instances. However, composing a large image database is challenging for the provider of an image-based Captcha as each image needs to be pre-labeled manually. We see several options to get images labeled: (1) Volunteers can help label images. (2) Associated web services may be willing to embed label tasks in their processes. (3)Labeling images may be rewarded with micropayments; for example, platforms for human intelligence tasks, such as *Amazon Mechanical Turk*, may be used.

With regard to research, our study has the following limitations, which at the same time provide avenues for further investigation:

1. The usability results of our empirical study are based on Captcha instances with less than 25 images. Our security analysis shows that such instances are secure unless modern AI-based algorithms are used to perform automated gender recognition. While our security analysis reveals the relationship between the numbers of images shown, gender recognition rates achieved by AI-based algorithms and the security level of a *Farett-Gender* and *Farett-Gender&Age* instance, further empirical studies are needed to assess the usability of instances when images are altered in terms of number and/or quality.

2. We did not control for cultural issues. Different cultures may affect the users' willingness to distinguish pictures based on gender.

3. Our study and the findings are based on the particular designs of *Farett-Age* and *Farett-Age&Gender*. As we did not implement and test other designs and variations (parametrizations), such as a higher number of pictures, we do not know how usability would be affected.

Such modifications of the proposed Captchas may become necessary when methods of artificial intelligence improve and threaten to diminish the level of security provided by the Captchas.

4. Different facial attributes (cf. [52] for a comprehensive overview) could serve as a basis for designing, implementing and evaluating further face recognition Captchas.

5. As mentioned in Subsection 4.2, an attacker can use AI-based algorithms not only for determining the age of given faces and then selecting the youngest or oldest one, s/he can also try to order a given set of faces according to age and then select the youngest or oldest one based on comparisons. Future research need to analyze to what extent such attacks will affect the security of the proposed Farett Captchas.

6. In our empirical study, the proposed Captchas methods are compared with text-based Captcha methods only. Further empirical research is required to compare additional types of Captchas, including other image-based Captchas, with the suggested methods.

Acknowledgements

We would like to thank the chair of Prof. Dr. Dogan Kesdogan (University of Regensburg, Germany) for technically hosting our web-based study and Ms. Franziska Tensing for her editorial assistance. This research was supported by the “Bavarian State Ministry of Education, Science and the Arts” as part of the FORSEC research association. Portions of the research in this paper use the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office.

Appendix A. Literature search process

We searched the literature using two types of data sources. First, we conducted a title search in pertinent journal and conference databases, namely Business Premier Source, IEEE Explore and ACM Digital Library. We used the search string “((CAPTCHA OR HIP[19]) and (usability OR accessibility OR security)) OR (face-recognition and (gender OR age))” and did not

[19] HIP abbreviates *Human Interaction Proof.*

limit the search period. Second, we scanned the table of contents of journals and conference proceedings for the period 2000-2014. We considered ACM Transactions on Information and System Security (TISSEC), Computers & Security, IEEE Transactions on Information Forensics and Security, IEEE Security and Privacy, Journal of Information System Security (JISSEC), International Journal of Information Security and Privacy (IJISP), Journal of Computer Security and Journal of Information Security and Privacy.

Appendix B. Proof of equation (4)

Without loss of generality, we assume that an attacker has to identify all k (out of n images) that show a female person. We further assume that the probability of correct gender classification of each image is p. Let $E_{ij}, i = 0, \ldots, k; j = 0, \ldots, (n-k)$, denote the event that the (AI-based algorithm of an) attacker correctly classifies exactly i of k female persons and correctly classifies exactly j of $(n-k)$ male persons. The E_{ij} events partition the set of all possible events so that the success probabilities for each of the E_{ij} events add to the overall success probability $\theta'_{Farett-Gender}$ of an attacker. We now determine the success probabilities for all E_{ij} in increasing order of i.

Case $i = 0$: The probability of the occurrence of $E_{0(n-k)}$ equals $(1-p)^k \cdot p^{n-k}$. In this case, all images are classified as male so that the attacker has to select k out of n images at random. From the attacker's perspective, all of the $\binom{n}{k}$ options are equally promising but only one option is the correct one. Overall, the probability with which the attacker succeeds under the event $E_{0(n-k)}$ equals

$$(1-p)^k \cdot p^{n-k} \cdot \frac{1}{\binom{n}{k}}. \tag{B.1}$$

In each event $E_{0j}, 0 \leq j < (n-k)$, the attacker misclassifies at least one male person as female and, at the same time, does not correctly classify any of the female persons. Thus, the attacker chooses at least one of the misclassified male persons and must finally fail. As a result, the success probability amounts to zero.

Case $i = 1$: The probability of the occurrence of $E_{1(n-k)}$ equals $(1-p)^{k-1} \cdot p^{n-k+1} \cdot \binom{k}{1}$: There are $\binom{k}{1}$ options to select exactly one female person which is correctly classified, all other images are classified as male. The correctly classified female person is selected by the attacker. Then the attacker has to additionally select $(k-1)$ out of $(n-1)$ images at random. From the attacker's perspective, again, all of the $\binom{n-1}{k-1}$ options are equally promising but only one option is correct. Overall, the probability with which the attacker succeeds under the event $E_{1(n-k)}$ equals

$$\binom{k}{1} \cdot (1-p)^{k-1} \cdot p^{n-k+1} \cdot \frac{1}{\binom{n-1}{k-1}}. \tag{B.2}$$

Analogously to case $i = 0$, in each event $E_{1j}, 0 \leq j < (n-k)$, the attacker misclassifies at least one male person as female and, at the same time, correctly classifies only one of the female persons. Thus, the attacker chooses at least one of the misclassified male persons and must fail (the success probability equals zero).

Cases $\boldsymbol{i = 2, \ldots, (k-1)}$: These cases can be analyzed analogously to the previous two cases. The success probability under the event $E_{i(n-k)}$ equals

$$\binom{k}{i} \cdot (1-p)^{k-i} \cdot p^{n-k+i} \cdot \frac{1}{\binom{n-i}{k-i}}. \tag{B.3}$$

To simplify equation (B.3), we use the following transformation:

$$\begin{aligned}
\binom{k}{i} \cdot \frac{1}{\binom{n-i}{k-i}} &= \frac{k!}{i! \cdot (k-i)!} \cdot \frac{1}{\frac{(n-i)!}{(k-i)! \cdot (n-i-[k-i])!}} \\
&= \frac{k! \cdot (n-k)!}{i! \cdot (n-i)!} \\
&= \frac{k! \cdot (n-k)!}{n!} \cdot \frac{n!}{i! \cdot (n-i)!} \\
&= \frac{1}{\binom{n}{k}} \cdot \binom{n}{i}.
\end{aligned} \tag{B.4}$$

Using equation (B.4) in equation (B.3), we finally yield

$$\frac{1}{\binom{n}{k}} \cdot (1-p)^{k-i} \cdot p^{n-k+i} \cdot \binom{n}{i}. \tag{B.5}$$

Case $\boldsymbol{i = k}$: This case differs from the previous ones as the misclassification of male persons now does not necessarily lead to the failure of the attacker due to the correct classification of all female persons. We distinguish the subcases $j = 0, \ldots, n-k$.

Subcase $\boldsymbol{j = 0}$: In this subcase, all female persons are correctly classified and all male persons are misclassified. The probability of this subcase equals $p^k \cdot (1-p)^{n-k}$. Now the attacker has to choose k images out of n images. From the attacker's perspective, all of the $\binom{n}{k}$ options are equally promising but only one option is the correct one. Overall, the probability with which the attacker succeeds under the event E_{k0} equals

$$p^k \cdot (1-p)^{n-k} \cdot \frac{1}{\binom{n}{k}}. \tag{B.6}$$

Subcase $j = 1$: In this subcase, all female persons and only one male person are classified correctly; there are $\binom{n-k}{1}$ options to correctly classify exactly one male person. Now the attacker has to choose k out of $n-1$ options. Similarly to subcase E_{k0}, from the attacker's perspective, all of the $\binom{n-1}{k}$ options are equally promising but only one option is the correct one. Overall, the probability with which the attacker succeeds under the event E_{k1} equals

$$\binom{n-k}{1} \cdot p^k \cdot (1-p)^{n-k-1} \cdot p \cdot \frac{1}{\binom{n-1}{k}} = \binom{n-k}{1} \cdot p^{k+1} \cdot (1-p)^{n-k-1} \cdot \frac{1}{\binom{n-1}{k}}. \tag{B.7}$$

Subcases $j = 2, \ldots, (n-k)$: These cases can be analyzed analogously to the previous two cases, resulting in the success probability (event E_{kj})

$$\binom{n-k}{j} \cdot (1-p)^{n-k-j} \cdot p^{k+j} \cdot \frac{1}{\binom{n-j}{k}}. \tag{B.8}$$

Analogously to using equation (B.4), we yield

$$\frac{1}{\binom{n}{k}} \cdot (1-p)^{n-k-j} \cdot p^{k+j} \cdot \binom{n}{j}. \tag{B.9}$$

Based on equations (B.5) and (B.9), the overall success probability of an attacker amounts to

$$
\begin{aligned}
\theta'_{Farett-Gender} &= \sum_{i=0}^{k-1} \frac{1}{\binom{n}{k}} \cdot (1-p)^{k-i} \cdot p^{n-k+i} \cdot \binom{n}{i} + \\
&\quad \sum_{j=0}^{n-k} \frac{1}{\binom{n}{k}} \cdot (1-p)^{n-k-j} \cdot p^{k+j} \cdot \binom{n}{j} \\
&= \frac{1}{\binom{n}{k}} \cdot \left[\sum_{i=0}^{k-1} (1-p)^{k-i} \cdot p^{n-k+i} \cdot \binom{n}{i} + \right. \\
&\quad \left. \sum_{j=0}^{n-k} (1-p)^{n-k-j} \cdot p^{k+j} \cdot \binom{n}{j} \right] \quad \square \qquad \text{(B.10)}
\end{aligned}
$$

References

[1] Ahmad, A. E., Yan, J., Tayara, M., 2011. The Robustness of Google CAPTCHA's. Tech. Rep. September, School of Computer Science, Newcastle University, UK.
URL http://homepages.cs.ncl.ac.uk/jeff.yan/google.pdf

[2] Ahn, L. V., Maurer, B., McMillen, C., Abraham, D., Blum, M., 2008. reCAPTCHA: Human-based character recognition via web security measures. Science 321 (5895), 1465–1468.
URL http://www.sciencemag.org/content/321/5895/1465.short

[3] Al-Khalifa, H. S., 2014. An empirical pilot study of captcha complexity using eye tracking. In: Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services. ACM, pp. 175–179.

[4] Ali, H., Tariq, U. U., Abid, M., 2014. Learning discriminating features for gender recognition of real world faces. International Journal of Image and Graphics 14 (03), 1450011.

[5] Baecher, P., Büscher, N., Fischlin, M., Milde, B., 2011. Breaking recaptcha: a holistic approach via shape recognition. In: Future Challenges in Security and Privacy for Academia and Industry. Springer, pp. 56–67.

[6] Baird, H. S., Bentley, J. L., 2005. Implicit CAPTCHAs. In: Proceedings of the IS&T/SPIE Document Recognition & Retrieval XII Conference. pp. 1–6.

[7] Basso, A., Bergadano, F., 2010. Anti-bot Strategies Based On Human Interactive Proofs. In: Handbook of Information and Communication Security. Springer, pp. 273–291.

[8] Basso, A., Sicco, S., 2009. Preventing massive automated access to web resources. Computers & Security 28 (3), 174–188.

[9] Bhattarai, B., Mignon, A., Jurie, F., Furon, T., 2014. Puzzling face verification algorithms for privacy protection. In: IEEE Workshop on Information Forensics and Security. IEEE, pp. 66 – 71.

[10] Bigham, J., Cavender, A., 2009. Evaluating existing audio CAPTCHAs and an interface optimized for non-visual use. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1829–1838.

[11] Brooke, J., 1996. Usability evaluation in industry. CRC Press, Ch. SUS: 'A quick and dirty' usability scale, pp. 189–194.

[12] Bursztein, E., Aigrain, J., Moscicki, A., Mitchell, J. C., 2014. The end is nigh: Generic solving of text-based captchas. In: 8th USENIX Workshop on Offensive Technologies. pp. 1–15.

[13] Bursztein, E., Bethard, S., Fabry, C., Mitchell, J., Jurafsky, D., 2010. How good are humans at solving captchas? a large scale evaluation. In: 2010 IEEE Symposium on Security and Privacy. pp. 399–413.

[14] Bursztein, E., Martin, M., Mitchell, J., 2011. Text-based CAPTCHA strengths and weaknesses. Proceedings of the 18th ACM conference on Com-

puter and communications security - CCS '11, 125.
URL http://dl.acm.org/citation.cfm?doid=2046707.2046724

[15] Bursztein, E., Moscicki, A., Fabry, C., Bethard, S., Mitchell, J. C., Jurafsky, D., 2014. Easy does it: more usable captchas. In: Proceedings of the 32nd annual ACM conference on Human factors in computing systems. ACM, pp. 2637–2646.

[16] Carnegie Mellon University, 2010. Captcha: Telling humans and computers apart automatically. http://www.captcha.net.

[17] Castrillón-Santana, M., Lorenzo-Navarro, J., Ramón-Balmaseda, E., 2015. On using periocular biometric for gender classification in the wild. Pattern Recognition Letters.

[18] Chandavale, A. A., Sapkal, A. M., Jalnekar, R. M., 2009. Algorithm to break visual captcha. In: 2nd International Conference on Emerging Trends in Engineering and Technology. IEEE, pp. 258–262.

[19] Chang, H. H., Chen, S. W., Oct. 2009. Consumer perception of interface quality, security, and loyalty in electronic commerce. Information & Management 46 (7), 411–417.
URL http://linkinghub.elsevier.com/retrieve/pii/S0378720609000901

[20] Chellapilla, K., Larson, K., Simard, P., Czerwinski, M., 2005. Building segmentation based human-friendly human interaction proofs (HIPs). Human Interactive Proofs, 173–185.
URL http://www.springerlink.com/index/h9kxwalaguwy5qyd.pdf

[21] Chellapilla, K., Larson, K., Simard, P. Y., Czerwinski, M., 2005. Computers beat humans at single character recognition in reading based human interaction proofs (hips). In: Proceedings of the 2nd Conference on Email and Anti-Spam. Palo Alto, pp. 21–22.

[22] Chellapilla, K., Simard, P., 2004. Using machine learning to break visual human interaction proofs (hips). Advances in neural information processing systems 17, 265–272.

[23] Confident Technologies, 2015. Confident CAPTCHA. http://confidenttechnologies.com/confident-captcha/.

[24] Conti, M., Guarisco, C., Spolaor, R., 2015. Captchastar! a novel captcha based on interactive shape discovery. arXiv preprint arXiv:1503.00561.

[25] Cruz-Perez, C., Starostenko, O., Uceda-Ponga, F., Alarcon-Aquino, V., Reyes-Cabrera, L., 2012. Breaking recaptchas with unpredictable collapse: heuristic character segmentation and recognition. In: Pattern Recognition. Springer, pp. 155–165.

[26] Dago-Casas, P., González-Jiménez, D., Yu, L. L., Alba-Castro, J. L., 2011. Single-and cross-database benchmarks for gender classification under unconstrained settings. In: 2011 IEEE International Conference on Computer Vision Workshops. pp. 2152–2159.

[27] Datta, R., Li, J., Wang, J., 2009. Exploiting the Human–Machine Gap in Image Recognition for Designing CAPTCHAs. IEEE Transactions onInformation Forensics and Security 4, 504–518.
URL `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4956990`

[28] Driessen, B., Dürmuth, M., 2013. Achieving anonymity against major face recognition algorithms. In: Communications and Multimedia Security. Springer, pp. 18–33.

[29] El Ahmad, A., Yan, J., Tayara, M., 2011. The Robustness of Google CAPTCHA's. Computing Science, Newcastle University.

[30] El Shafey, L., Khoury, E., Marcel, S., 2014. Audio-visual gender recognition in uncontrolled environment using variability modeling techniques. In: 2014 IEEE International Joint Conference on Biometrics. IEEE, pp. 1–8.

[31] Elson, J., Douceur, J., Howell, J., Saul, J., 2007. Asirra: a CAPTCHA that exploits interest-aligned manual image categorization. CCS 7, 366–374.
URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.187.8153&rep=rep1&type=pdf`

[32] Fang, K., Bu, Z., Xia, Z. Y., 2012. Segmentation of captchas based on complex networks. In: Artificial Intelligence and Computational Intelligence. Springer, pp. 735–743.

[33] Fidas, C., Hussmann, H., Belk, M., Samaras, G., 2015. ihip: Towards a user centric individual human interaction proof framework. In: Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems. ACM, pp. 2235–2240.

[34] Fidas, C. A., Voyiatzis, A. G., Avouris, N. M., 2011. On the necessity of user-friendly CAPTCHA. Proceedings of the 2011 annual conference on Human factors in computing systems, 2623.
URL `http://dl.acm.org/citation.cfm?doid=1978942.1979325`

[35] Fritsch, C., Netter, M., Reisser, A., Pernul, G., 2010. Attacking Image Recognition Captchas A Naive but Effective Approach. In: Katsikas, S., Lopez, J., Soriano, M. (Eds.), Proceedings of the 7th International Conference on Trust, Privacy & Security in Digital Business. Vol. 6264 of LNCS. Springer, Berlin, Heidelberg, pp. 13–25.

[36] Fu, Y., Guo, G., Huang, T. S., 2010. Age synthesis and estimation via faces: A survey. Pattern Analysis and Machine Intelligence, IEEE Transactions on 32 (11), 1955–1976.

[37] Gao, H., Wang, W., Qi, J., Wang, X., Liu, X., Yan, J., 2013. The robustness of hollow captchas. In: Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security. ACM, pp. 1075–1086.

[38] Gao, W., Ai, H., 2009. Face gender classification on consumer images in a multiethnic environment. Advances in Biometrics, 169–178.
URL `http://www.springerlink.com/index/wjr27w0564048036.pdf`

[39] Geng, X., Smith-Miles, K., 2009. Facial age estimation by multilinear

subspace analysis. IEEE International Conference on Acoustics, Speech and Signal Processing, 865–868.
URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4959721

[40] Golle, P., 2008. Machine learning attacks against the Asirra CAPTCHA. Proceedings of the 15th ACM conference on Computer and communications security, 535–542.
URL http://dl.acm.org/citation.cfm?id=1455838

[41] Google, 2013. The Official CAPTCHA Site.
URL http://www.google.com/recaptcha/faq

[42] Gossweiler, R., Kamvar, M., Baluja, S., 2009. What's up CAPTCHA?: a CAPTCHA based on image orientation. Proceedings of the 18th international conference on World wide web, 841–850.
URL http://dl.acm.org/citation.cfm?id=1526822

[43] Goswami, G., Singh, R., Vatsa, M., Powell, B., Noore, A., 2012. Face recognition captcha. In: Proceedings of the Fifth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS). IEEE, pp. 412–417.

[44] Hidalgo, J. M. G., Alvarez, G., 2011. captchas: An Artificial Intelligence Application To Web Security. Advances in Computers 83, 109–181.

[45] Horng, W.-B., Lee, C.-P., Chen, C.-W., 2001. Classification of age groups based on facial features. Tamkang Journal of Science and Engineering 4 (3), 183–192.

[46] Houck, C., 2010. Decoding recaptcha. Defcon 18, Las Vegas,http://n3on.org/projects/reCAPTCHA/.

[47] Huang, G. B., Learned-Miller, E., 2014. Labeled faces in the wild: Updates and new reporting procedures. Dept. Comput. Sci., Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep, 14–003.

[48] Huang, G. B., Ramesh, M., Berg, T., Learned-Miller, E., 2007. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. rep., Technical Report 07-49, University of Massachusetts, Amherst.

[49] Kim, J., Kim, S., Yang, J., Ryu, J.-h., Wohn, K., 2013. Facecaptcha: a captcha that identifies the gender of face images unrecognized by existing gender classifiers. Multimedia Tools and Applications, 1–23.

[50] Kluever, K. A., Zanibbi, R., 2009. Balancing usability and security in a video captcha. In: Proceedings of the 5th Symposium on Usable Privacy and Security. ACM, p. 14.

[51] Korayem, M., Mohamed, A., Crandall, D., Yampolskiy, R., 2012. Solving Avatar Captchas Automatically. cs.indiana.edu.
URL https://www.cs.indiana.edu/~mkorayem/paper/CAPTCHA.pdf

[52] Kumar, N., 2011. Attribute and simile classifiers for face verification.
URL http://neerajkumar.org/projects/faceverification/

[53] Lupkowski, P., Urbanski, M., 2008. SemCAPTCHA user-friendly alternative for OCR-based CAPTCHA systems. Computer Science and Information Technology, 325–329.

[54] Ma, J., Badaoui, B., Chamoun, E., 2009. A generalized method to solve text-based captchas.

[55] Mäkinen, E., Raisamo, R., Jul. 2008. An experimental comparison of gender classification methods. Pattern Recognition Letters 29 (10), 1544–1556.
URL `http://linkinghub.elsevier.com/retrieve/pii/S0167865508001116`

[56] Makris, C., Town, C., 2014. Character segmentation for automatic captcha solving. Open Computer Science Journal 1 (1).

[57] Mehra, M., Agarwal, M., Pawar, R., Shah, D., 2011. Mitigating denial of service attack using CAPTCHA mechanism. In: Proceedings of the International Conference & Workshop on Emerging Trends in Technology - ICWET '11. ACM Press, New York, New York, USA, p. 284.
URL `http://portal.acm.org/citation.cfm?doid=1980022.1980086`

[58] Meutzner, H., Gupta, S., Kolossa, D., 2015. Constructing secure audio captchas by exploiting differences between humans and machines. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. ACM, pp. 2335–2338.

[59] Meutzner, H., Nguyen, V.-H., Holz, T., Kolossa, D., 2014. Using automatic speech recognition for attacking acoustic captchas: the trade-off between usability and security. In: Proceedings of the 30th Annual Computer Security Applications Conference. ACM, pp. 276–285.

[60] Mitchell, M. L., Jolley, J. M., 2012. Research design explained, 8th Edition. Cengage Learning.

[61] Mohamed, M., Sachdeva, N., Georgescu, M., Gao, S., Saxena, N., Zhang, C., Kumaraguru, P., van Oorschot, P. C., Chen, W.-B., 2014. A three-way investigation of a game-captcha: automated attacks, relay attacks and usability. In: Proceedings of the 9th ACM symposium on Information, computer and communications security. ACM, pp. 195–206.

[62] Morein, W., Stavrou, A., Cook, D., Keromytis, A., Misra, V., Rubenstein, D., 2003. Using graphic turing tests to counter automated ddos attacks against web servers. In: Proceedings of the 10th ACM conference on Computer and communications security. ACM, pp. 8–19.

[63] Moy, G., Jones, N., Harkless, C., Potter, R., 2004. Distortion estimation techniques in solving visual captchas. In: Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Vol. 2. IEEE, pp. II–23.

[64] Namprempre, C., Dailey, M. N., 2007. Mitigating dictionary attacks with text-graphics character captchas. IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences 90 (1), 179–186.

[65] Nguyen, V. D., Chow, Y.-W., Susilo, W., 2014. On the security of text-based

3d captchas. Computers & Security 45, 84–99.

[66] Nielsen, J., 1993. Usability engineering. Academic Press, London.

[67] Nielsen, J., 2012. Usability 101: Introduction to usability. http://www.useit.com/alertbox/20030825.html.

[68] Olalere, A., Feng, J. H., Lazar, J., Brooks, T., 2014. Investigating the effects of sound masking on the use of audio captchas. Behaviour & Information Technology 33 (9), 919–928.

[69] Pashalidis, A., Mavrogiannopoulos, N., Ferrer, X., Bermejo Olaizola, B., 2012. For human eyes only: security and usability evaluation. In: Proceedings of the 2012 ACM workshop on Privacy in the electronic society. ACM, pp. 129–140.

[70] Penninger, S., Meier, S., Federrath, H., 2012. Usability von CAPTCHA-Systemen. In: Waidner, M., Neeraj, S. (Eds.), Proceedings of Sicherheit 2012 (Sicherheit, Schutz und Zuverlässigkeit). Vol. P-195 of Lecture Notes in Informatics. Kőllen-Verlag, Bonn, pp. 199–208.
URL `http://epub.uni-regensburg.de/23565`

[71] Phillips, P. J., Moon, H., Rizvi, S., Rauss, P. J., et al., 2000. The feret evaluation methodology for face-recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (10), 1090–1104.

[72] Phillips, P. J., Wechsler, H., Huang, J., Rauss, P. J., 1998. The feret database and evaluation procedure for face-recognition algorithms. Image and Vision Computing 16 (5), 295–306.

[73] Pinkas, B., Sander, T., 2002. Securing passwords against dictionary attacks. Proceedings of the 9th ACM conference on Computer and communications security - CCS '02, 161.
URL `http://portal.acm.org/citation.cfm?doid=586110.586133`

[74] Roshanbin, N., Miller, J., 2013. A survey and analysis of current captcha approaches. Journal of Web Engineering 12 (1), 1–40.
URL `http://dl.acm.org/citation.cfm?id=2481563`

[75] Rui, Y., Liu, Z., Jun. 2004. ARTiFACIAL: Automated Reverse Turing test using FACIAL features. Multimedia Systems 9 (6), 493–502.
URL `http://www.springerlink.com/openurl.asp?genre=article&id=doi:10.1007/s00530-003-0122-3`

[76] Sauer, G., Hochheiser, H., 2008. Towards a universally usable CAPTCHA. Proceedings of the 4th Symposium on Usable Privacy and Security, 2–5.
URL `http://cups.cs.cmu.edu/soups/2008/SOAPS/sauer.pdf`

[77] Sergei, K., n.d. Kcaptcha project. http://www.captcha.ru/en/kcaptcha/.

[78] Shakhnarovich, G., Viola, P., Moghaddam, B., et al., 2002. A unified learning framework for real time face detection and classification. In: Proceedings of the Fifth IEEE International Conference on Automatic Face and Gesture Recognition. pp. 14–21.

[79] Shan, C., Mar. 2012. Learning local binary patterns for gender classification

on real-world face images. Pattern Recognition Letters 33 (4), 431–437.
URL http://linkinghub.elsevier.com/retrieve/pii/S0167865511001607

[80] Sheskin, D. J., 2003. Handbook of parametric and nonparametric statistical procedures. crc Press.

[81] Shirali-Shahreza, S., Penn, G., Balakrishnan, R., Ganjali, Y., 2013. Seesay and hearsay captcha for mobile interaction. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, pp. 2147–2156.

[82] Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., Jain, R., 2000. Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence 22 (12), 1349–1380.

[83] Starostenko, O., Cruz-Perez, C., Uceda-Ponga, F., Alarcon-Aquino, V., 2015. Breaking text-based captchas with variable word and character orientation. Pattern Recognition 48 (4), 1101–1112.

[84] Sutherland, C., 2012. Usability and Security of Text-based CAPTCHAs. wiki.umn.edu.
URL https://wiki.umn.edu/pub/UmmCSciSeniorSeminar/Spring2012Talks/CodySutherland.pdf

[85] Tam, J., Simsa, J., Sean, H., von Ahn, L., 2008. Breaking audio captchas. Advances in Neural Information Processing Systems 1, 1–8.
URL http://accounts-service-login.com/nph-g.cgi/00/http/www.captcha.net/Breaking_Audio_CAPTCHAs.pdf

[86] Thulin, M., Masek, P., n/a. Software Quality Evaluation of Face Recognition APIs & Libraries.

[87] University of Massachusetts Amherst, 2015. Labeled faces in the wild.
URL http://vis-www.cs.umass.edu/lfw/

[88] Viola, P., Jones, M., 2004. Robust real-time face detection. International journal of computer vision 57, 137–154.
URL http://www.springerlink.com/index/q70v4h6715v5p152.pdf

[89] von Ahn, L., Blum, M., Hopper, N., Langford, J., 2003. Captcha: Using hard ai problems for security. In: Biham, E. (Ed.), Advances in Cryptology EUROCRYPT 2003. Vol. 2656 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 294–311.
URL http://dx.doi.org/10.1007/3-540-39200-9_18

[90] von Ahn, L., Blum, M., Langford, J., 2002. Telling humans and computers apart automatically or how lazy cryptographers do ai. Computer Science Department, 149.

[91] von Ahn, L., Blum, M., Langford, J., 2002. Telling humans and computers apart automatically or how lazy cryptographers do ai. Computer Science Department, 149.

[92] von Ahn, L., Blum, M., Langford, J., Feb. 2004. Telling humans and computers apart automatically. Communications of the ACM 47 (2), 56–60.

URL http://doi.acm.org/10.1145/966389.966390

[93] von Ahn, L., Maurer, B., McMillen, C., Abraham, D., Blum, M., 2008. recaptcha: Human-based character recognition via web security measures. Science 321 (5895), 1465–1468.

[94] Wilkins, J., 2009. Strong CAPTCHA Guidelines v1.2. Retrieved Nov 10, 1–18.

[95] Xu, Y., Reynaga, G., Chiasson, S., Frahm, J.-M., Monrose, F., van Oorschot, P. C., 2012. Security and Usability Challenges of Moving-Object CAPTCHAs: Decoding Codewords in Motion. In: USENIX Security Symposium. pp. 49–64.

[96] Yan, J., Ahmad, A. E., 2007. Breaking visual captchas with naive pattern recognition algorithms. Twenty-Third Annual Computer Security Applications Conference.
URL http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4412996

[97] Yan, J., El Ahmad, A. S., 2008. A low-cost attack on a Microsoft captcha. Proceedings of the 15th ACM conference on Computer and communications security - CCS '08, 543.
URL http://portal.acm.org/citation.cfm?doid=1455770.1455839

[98] Yan, J., El Ahmad, A. S., 2008. Usability of CAPTCHAs or usability issues in CAPTCHA design. Proceedings of the 4th symposium on Usable privacy and security - SOUPS '08, 44.
URL http://portal.acm.org/citation.cfm?doid=1408664.1408671

[99] Zelkowitz, M., 2011. Security on the Web. Elsevier.
URL http://books.google.com/books?hl=de&lr=&id=L904A5qzeioC&pgis=1

[100] Zhu, B. B., Yan, J., Li, Q., Yang, C., Liu, J., Xu, N., Yi, M., Cai, K., 2010. Attacks and design of image recognition CAPTCHAs. Proceedings of the 17th ACM conference on Computer and communications security - CCS '10, 187.
URL http://portal.acm.org/citation.cfm?doid=1866307.1866329

List of Tables

Captcha		Usability	Security
Text-based	Google (reCAPTCHA)	[13]^{1,2}, [70]^{1,2,3,4}, [98]^{1,2,3}	[1]^{5}, [2], [5]^{7}, [14]^{5,7}, [20]^{5}, [21]^{5}, [22]^{5}, [84]^{5}, [94]^{5,7}
	Microsoft	[13]^{1,2}, [98]^{1,2,3}	[14]^{5,7}, [20]^{5}, [21]^{5}, [22]^{5}, [84]^{5},[97]^{5}
	Yahoo	[13]^{1,2}, [98]^{1,2,3}	[14]^{5,7}, [20]^{5}, [21]^{5}, [22]^{5}, [84]^{5}
	Others	[3]^{2}, [13]^{1,2}, [15]^{1,2}, [69]^{1,2}, [81]^{1,2,4}	[12]^{7}, [14]^{5,7}, [18]^{7}, [20]^{5}, [21]^{5}, [22]^{5}, [25]^{7}, [32]^{7}, [37]^{7}, [54]^{7}, [56]^{7}, [63]^{5}, [65]^{7}, [69]^{7}, [74]^{5,7}, [83]^{7}, [84]^{5}, [94]^{5,7}, [96]^{5}
Image-based	Microsoft (ASIRRA)	[70]^{1,2,3,4}	[31]^{5,6}, [35]^{5,7}, [40]^{5}, [100]^{5,7}
	IMAGINATION	[27]^{1}	[27]^{5}, [100]^{5,7}
	ARTiFACIAL		[75]^{5}, [100]^{5,7}
	HumanAuth		[35]^{5,7}
	Avatar Captcha		[51]^{5}
	Others	[8]^{1,2,3}, [43]^{1}, [49]^{1,2}	[8]^{5}
Audio-based	Google (reCAPTCHA audio)	[10]^{1,2,4},[13]^{1,2}, [59]^{1}, [76]^{1}	[59]^{7}, [85]^{5}
	Microsoft	[13]^{1,2}	
	Yahoo	[13]^{1,2}	
	Others	[10]^{1,2,4}, [13]^{1,2}, [58]^{1}, [59]^{1}, [68], [70]^{1,2,3,4}, [81]^{1,2,4}	[58]^{7}, [59]^{7}, [85]^{5}
Video-based		[50]^{1,2,4}, [95]^{1,2,3,4}	[50]*,[74]^{5,7}, [95]^{5}
Interaction-based		[42]^{1,4}, [61]^{1,2,4}	[42]^{5}, [61]^{7}

1 = Effectiveness, 2 = Efficiency, 3 = Learnability, 4 = Satisfaction, 5 = Media security
6 = Script security, 7 = Algorithmic/Randomness security, 8 = Security after wide-spread adoption
* = Paper does not address particular security requirements of video-based Captchas

Table B.1: Literature findings for usability and security of Captchas

Probability of correct gender detection p=80%		from n images				
		20	30	40	50	60
select k females	5	2.54	0.54	0.14	0.04	0.01
	7	1.99	0.33	0.06	0.01	<0.01
	9	1.81	0.25	0.04	<0.01	<0.01
	11	1.81	0.22	0.03	<0.01	<0.01

Table B.2: Probabilities (in %) with which an AI-based attack on *Farett-Gender* succeeds (p=80%)

Probability of correct gender detection p=85%		from n images				
		20	30	40	50	60
select k females	5	6.64	1.98	0.66	0.25	0.1
	7	5.68	1.45	0.4	0.12	0.04
	9	5.34	1.29	0.31	0.08	0.02
	11	5.34	1.12	0.26	0.06	0.02

Table B.3: Probabilities (in %) with which an AI-based attack on *Farett-Gender* succeeds (p=85%)

Probability of correct gender detection p=90%		from n images				
		20	30	40	50	60
select k females	5	16.88	7.36	3.35	1.59	0.79
	7	15.43	6.21	2.59	1.11	0.49
	9	14.89	5.67	2.24	0.91	0.37
	11	14.89	5.39	2.08	0.8	0.32

Table B.4: Probabilities (in %) with which an AI-based attack on *Farett-Gender* succeeds (p=90%)

p	n						
	16	20	24	28	32	36	40
80%	1.26	0.75	0.47	0.31	0.21	0.15	0.11
85%	2.42	1.60	1.09	0.76	0.54	0.39	0.29
90%	4.64	3.48	2.64	2.02	1.56	1.22	0.96

Table B.5: Probabilities (in %) with which an (gender: AI-based, age: bruteforce) attack on *Farett-Gender&Age* succeeds ($k = 3$)

p=q	n						
	16	20	24	28	32	36	40
80%	5.35	3.19	1.99	1.3	0.89	0.62	0.46
85%	11.33	7.49	5.08	3.53	2.52	1.83	1.37
90%	23.75	17.79	13.47	10.3	7.96	6.21	4.89

Table B.6: Probabilities (in %) with which an attack (gender: AI-based, age: AI) on *Farett-Gender&Age* ($k = 3$) succeeds

Age group (Count)	male	female
<20	2	1
20-29	60	18
30-39	24	2
40-50	9	0
>50	1	1
Total	96	22
Total count participants		**118**
Mean age		28.88
Standard deviation σ_{age}		7.03

Internet usage (in hours[10])	Count
<1	23
1-2	15
2-3	17
3-4	11
4-5	14
5-6	10
6-7	15
7-8	1
>8	12
PC-knowledge[11]	Count
0	1
1-2	1
3-4	0
5-6	9
7-8	48
9-10	58

Table B.7: Characteristics of participants

	Position			
	1	2	3	4
Group 1	A	B	D	C
Group 2	B	C	A	D
Group 3	C	D	B	A
Group 4	D	A	C	B

A: *reCAPTCHA*, B: *KCAPTCHA*
C: *Farett-Gender*, D: *Farett-Gender&Age*

Table B.8: 4x4 Latin Square design, based on [60, p. 542]

Usability		Captcha	
Criterion	Measure	**reCAPTCHA (rC)**	**KCAPTCHA (KC)**
Effectiveness	Error rate in %	Mean=0.11 (Median=0.0), SD=0.17 KC: ($V = 141.5, p < .001, N = 118$)	Mean=0.4 (Median=0.4), SD= 0.21
Efficiency	Solving time in s (overall)	Mean=11.37 (Median=9.87), SD=4.63 FG: ($V = 1044, p < .001, N = 118$) FG&A: ($V = 937, p < .001, N = 118$)	Mean=10.2 (Median=8.93), SD=4.38 rC: ($V = 5185, p < .001, N = 118$) FG: ($V = 712, p < .001, N = 118$) FG&A: ($V = 587.5, p < .001, N = 118$)
	Solving time in s (correctly solved)	Mean=11.47 (Median=9.99), SD=4.86 FG: ($V = 1239, p < .001, N = 118$) FG&A: ($V = 829, p < .001, N = 108$)	Mean=10.15 (Median=8.84), SD=7.14 rC: ($V = 5068, p < .001, N = 118$) FG: ($V = 729, p < .001, N = 118$) FG&A: ($V = 477, p < .001, N = 108$)
	Solving time in s (incorrectly solved)	Mean=10.57 (Median=9.23), SD=5.74 FG: ($V = 3, p = .04, N = 7$)	Mean=10.27 (Median=9.02), SD=4.57 rC: ($V = 579, p = .004, N = 39$) FG: ($V = 56, p < .001, N = 29$) FG&A: ($V = 224, p < .001, N = 59$)
Learnability	(ease+intuition)/2	Mean=7.42 (Median=7.5), SD=1.85 KC: ($V = 4932.5, p < .001, N = 118$)	Mean=5.8 (Median=6.0), SD=2.04
Satisfaction	Overall impression	Mean=6.66 (Median=7.0), SD=2.35 KC: ($V = 4834, p < .001, N = 118$)	Mean=4.8 (Median=5.0), SD=2.29

Usability		Captcha	
Criterion	Measure	**Farett-Gender (FG)**	**Farett-Gender&Age (FG&A)**
Effectiveness	Error rate in %	Mean=0.08 (Median=0.0), SD=0.16 KC: ($V = 5466, p < .001, N = 118$)	Mean=0.23 (Median=0.2), SD=0.31 KC: ($V = 3538, p < .001, N = 118$)
Efficiency	Solving time in s (overall)	Mean=14.46 (Median=13.42), SD=4.97 FG&A: ($V = 2551.5, p = .005, N = 118$)	Mean=15.41 (Median=13.98), SD=5.59
	Solving time in s (correctly solved)	Mean=14.05 (Median=13.10), SD=4.98 FG&A: ($V = 1620, p < .001, N = 108$)	Mean=15.08 (Median=14.02), SD=6.9
	Solving time in s (incorrectly solved)	Mean=19.03 (Median=17.35), SD=9.67	Mean=16.51 (Median=13.94), SD=8.31
Learnability	(ease+intuition)/2	Mean=7.93 (Median=8.5), SD= 2.06 rC: ($V = 2052.5, p = .007, N = 118$) KC: ($V = 950, p < .001, N = 118$) FG&A: ($V = 3724, p < .001, N = 118$)	Mean=7.29 (Median=8.0), SD=2.16 KC: ($V = 1325.5, p < .001, N = 118$)
Satisfaction	Overall impression	Mean=7.53 (Median=8.0), SD=2.43 rC: ($V = 1719, p = .003, N = 118$) KC: ($V = 797.5, p < .001, N = 118$)	Mean=6.99 (Median=8.0), SD=2.61 KC: ($V = 1107, p < .001, N = 118$)

rC: reCAPTCHA, KC: KCAPTCHA, FG: *Farett-Gender*, FG&A: *Farett-Gender&Age*

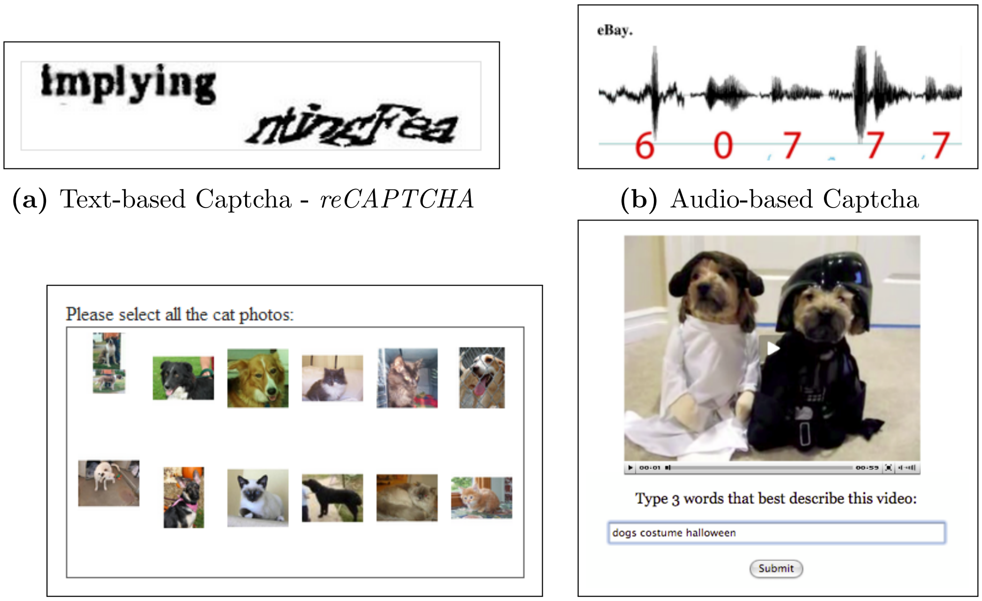
SD: standard deviation

Dominated procedure with level of significance (***: .001, **: .01), (value, reduced sample size of paired Wilcoxon rank sum test)

–: Wilcoxon rank sum test not applicable or no statistical significance at the .01 level

Table B.9: Results of the usability study

List of Figures



Please select all the cat photos:

Type 3 words that best describe this video:

dogs costume halloween

(a) Text-based Captcha - *reCAPTCHA*

(b) Audio-based Captcha

(c) Image-based Captcha - *ASIRRA*

(d) Video-based Captcha - *eBay*

Figure B.1: Examples of different types of Captchas

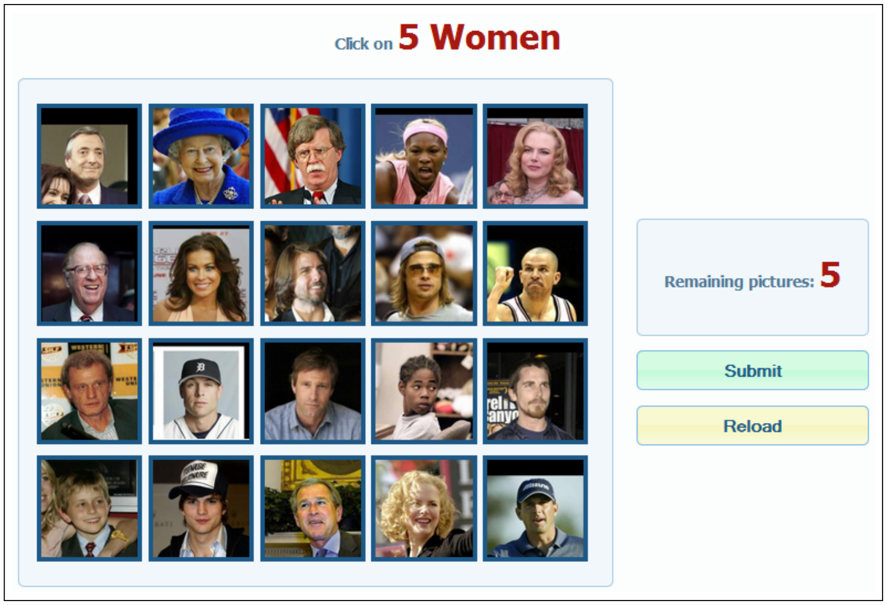
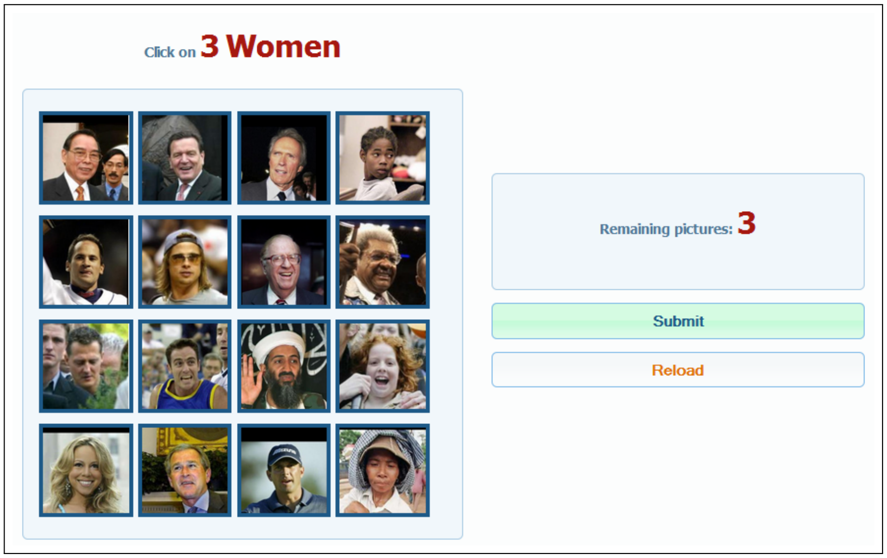


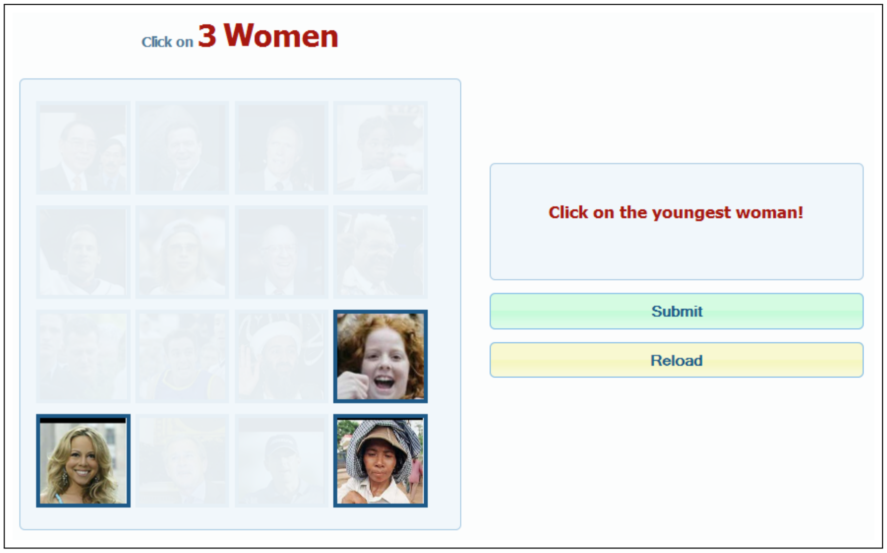
Figure B.2: Screenshot of a challenge of *Farett-Gender*



Click on 3 Women

Remaining pictures: 3

Figure B.3: Screenshot of a challenge of *Farett-Gender&Age* (task 1)



Click on 3 Women

Click on the youngest woman!

Figure B.4: Screenshot of a challenge of *Farett-Gender&Age* (task 2)

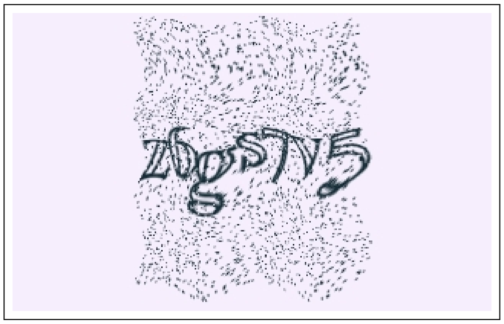
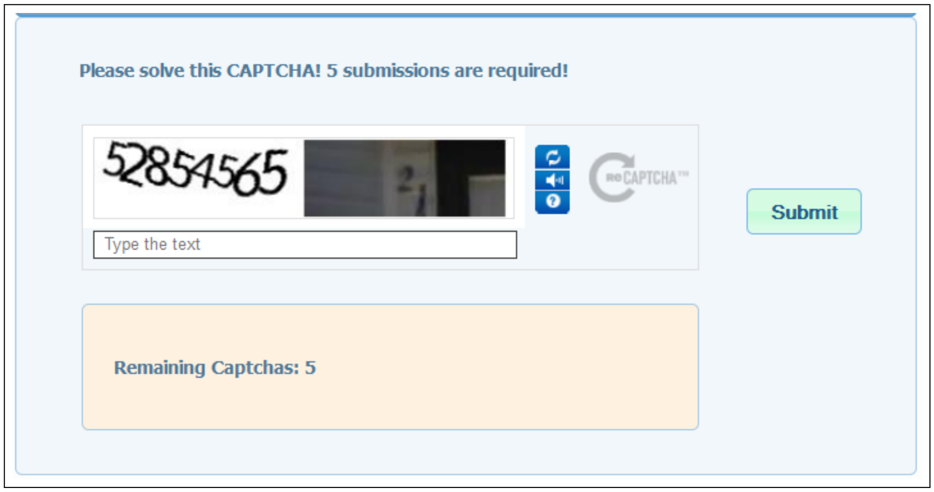


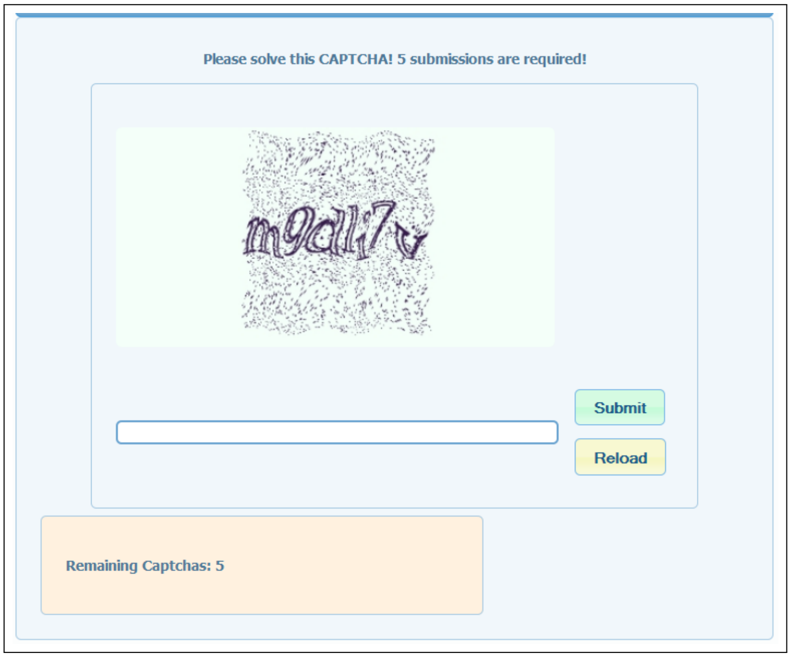
Figure B.5: *KCAPTCHA* - Example of a challenge



Please solve this CAPTCHA! 5 submissions are required!

Remaining Captchas: 5

Figure B.6: Screenshot of *reCAPTCHA*



Please solve this CAPTCHA! 5 submissions are required!

Remaining Captchas: 5

Figure B.7: Screenshot of *KCAPTCHA*

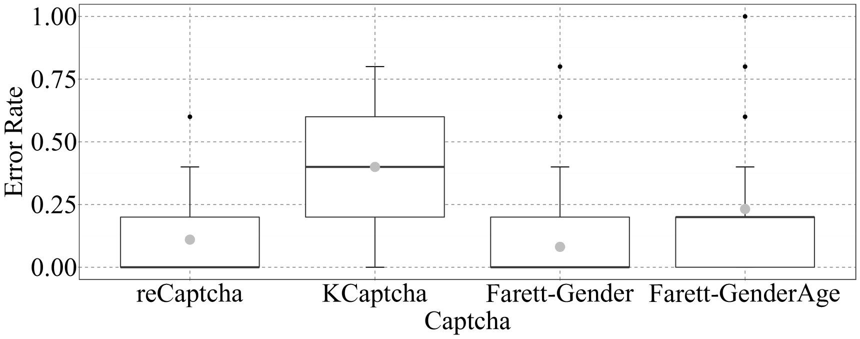


Figure B.8: Boxplots of error rates

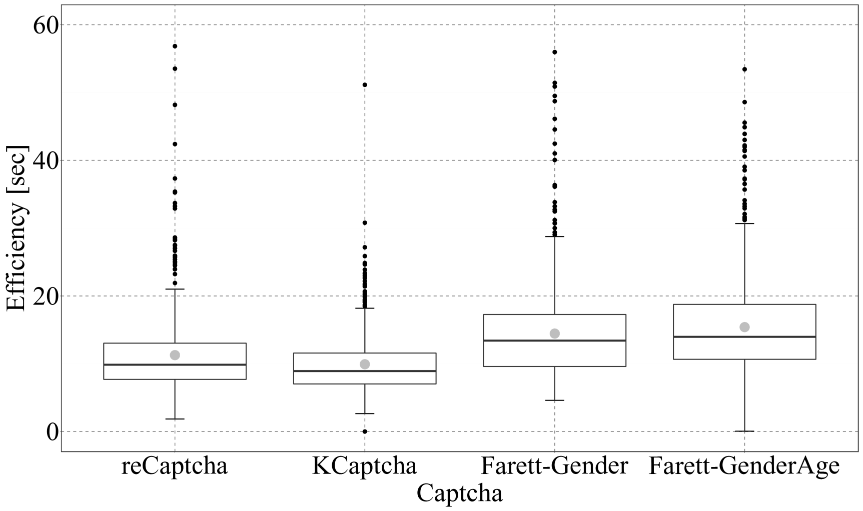


Figure B.9: Boxplots of task completion times

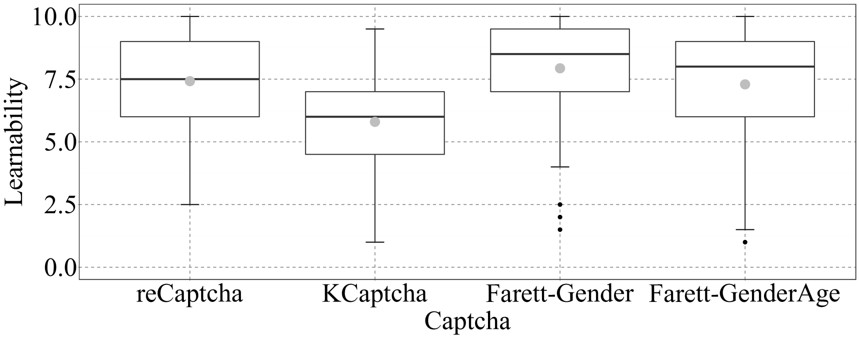


Figure B.10: Boxplots of learnability

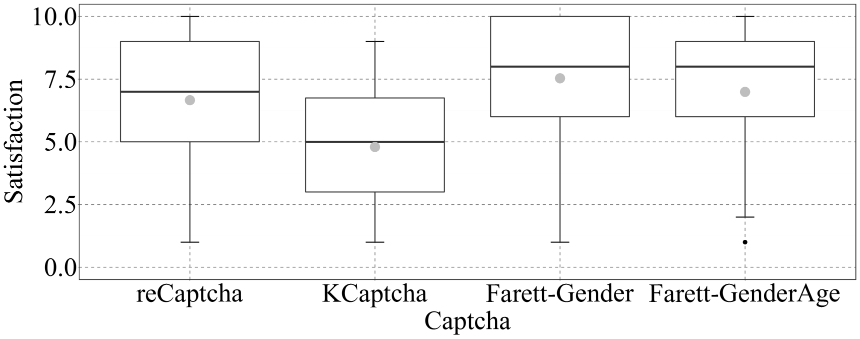


Figure B.11: Boxplots of satisfaction