Using artificially generated pictures in customer-facing systems: an evaluation study with data-driven personas

Joni Salminen , Soon-gyo Jung , Ahmed Mohamed Sayed Kamel , João M. Santos & Bernard J. Jansen

Published online: 06 Nov 2020.

Submit your article to this journal

Article views: 279

View related articles

View Crossmark data

Using artificially generated pictures in customer-facing systems: an evaluation study with data-driven personas

Joni Salminen[a,b], Soon-gyo Jung[a], Ahmed Mohamed Sayed Kamel[c], João M. Santos[d] and Bernard J. Jansen[a]

[a]Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar; [b]Turku School of Economics at the University of Turku, Turku, Finland; [c]Department of Clinical Pharmacy, Cairo University, Giza, Egypt; [d]Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal

ABSTRACT

We conduct two studies to evaluate the suitability of artificially generated facial pictures for use in a customer-facing system using data-driven personas. STUDY 1 investigates the quality of a sample of 1,000 artificially generated facial pictures. Obtaining 6,812 crowd judgments, we find that 90% of the images are rated medium quality or better. STUDY 2 examines the application of artificially generated facial pictures in data-driven personas using an experimental setting where the high-quality pictures are implemented in persona profiles. Based on 496 participants using 4 persona treatments (2×2 research design), findings of Bayesian analysis show that using the artificial pictures in persona profiles did not decrease the scores for Authenticity, Clarity, Empathy, and Willingness to Use of the data-driven personas.

1. Introduction

There is tremendous research interest concerning artificial image generation (AIG). The state-of-the-art studies in this field use *Generative Adversarial Networks* (GANs) (Goodfellow et al. 2014) and Conditional GANs (Lu, Tai, and Tang 2017) to generate images that are promised to be photorealistic and easily deployable. GANs have been applied, for example, to automatically create art (Tan et al. 2017), cartoons (Liu et al. 2018), medical images (Nie et al. 2017), and facial pictures (Karras, Laine, and Aila 2019), the latter including transformations such as increasing/decreasing a person's age or altering their gender (Antipov, Baccouche, and Dugelay 2017; Choi et al. 2018; Isola et al. 2017).

Due its low cost, AIG provides novel opportunities for a wide range of applications, including health-care (Nie et al. 2017), advertising (Neumann, Pyromallis, and Alexander 2018), and user analytics for human computer interaction (HCI) and design purposes (Salminen et al. 2019a). However, despite the far-reaching interest in AIG among academia and across industries, there is scant research on *evaluating the suitability of the generated images for practical use in deployed systems*. This means that the quality and impact of the artificial images on user perceptions are often neglected, lacking user studies of their deployment in real systems. This area of evaluation is an overlooked but critical area of research, as it is the 'final step' of deployment that actually determines if the quality of the AIG is good enough, as prior work has shown the impact that pictures can have on real systems (King, Lazard, and White 2020). Therefore, the impact of AIG on user experience (UX) and design applications is a largely unaddressed field of study, although with work in related areas of empathy (Weiss and Cohen 2019). For example, Weiss and Cohen (2019) that aspects of empathy with subjects in videos is complex in terms of encouraging or discouraging engagement with the content.

Most typically, artificial pictures are evaluated using *technical metrics* (Yuan et al. 2020) that are abstract and do not reflect user perceptions or UX. An example is the *Frèchet inception distance* (FID) (Heusel et al. 2017) that measures the similarity of two image distributions (i.e. the generated set and the training set). While metrics such as FID are without question necessary for measuring the *technical quality* of the generated images (Zhao et al. 2020), we argue there is also a substantial need for evaluating the *user experience* of the pictures for real-world systems and applications.

In this regard, the user study tradition from HCI is helpful – in addition to technical metrics, *user-centric metrics* gauging UX and user perceptions (Ashraf, Jaafar, and Sulaiman 2019; Brauner et al. 2019) can be deployed. The potential impact of AIG is

transformational, including domains of public relations, marketing, advertising, ecommerce sites, retail brochures, chatbots, virtual agents, design, and others. The use of artificially generated facial images is generally free of copyright restrictions and can allow for a wide range of demographic diversity (age, gender, ethnicity). Nonetheless, these benefits only hold if the pictures are 'good enough' for real applications. Given the multiple application areas of AIG, the results of an evaluation study measuring the impact of artificial facial pictures on UX is of immediate interest for researchers and practitioners alike.

To address the call for user studies concerning AIG, we carry out two evaluation studies: (a) one addressing the overall *perceived quality* of artificial pictures among crowd workers, and (b) another addressing user perceptions when implementing the pictures for data-driven personas (DDPs). Our research question is: *Are artificially generated facial pictures 'good enough' for a system requiring substantial images of people?*

DDPs are personas imaginary people representing real user segments, as defined traditionally in HCI (Cooper 2004) created from social media and Web analytics data (An et al. 2018a, 2018b). Although the DDP process may vary system to system, most will have the six major steps shown in Figure 1.

The advantage of DDPs, relative to traditional personas (Brangier and Bornet 2011) that are manually created and typically include 3–7 personas per set (Hong et al. 2018), is that one can create hundreds of DDPs from the data to reflect different behavioural and demographic nuances in the underlying user population (Salminen et al. 2018b). For example, a news organisation distributing its contents in social media platforms to audiences originating from dozens of countries can have dozens of audience segments relevant for different decision-making scenarios in other geographic areas (Salminen et al. 2018e). DDPs summarise these segments into easily approachable human profiles (see Figure 2 for example) that can be used within the organisation to understand the persona's needs (Nielsen 2019) and communicate (Salminen et al. 2018d) about these needs as a part of user-centric decision making (Idoughi, Seffah, and Kolski 2012).

One of the important issues for automatically creating DDPs from data is the availability of persona pictures – since DDP systems can create dozens of personas in near real time, there is a need for an inventory of pictures to use when rendering the personas for end users to view and interact with. Conceptually, this leads to a need for an AIG module that creates suitable persona pictures on demand (Salminen et al. 2019a). However, prior to engaging in system development, there is a need for ensuring that artificially generated pictures are not detrimental to user perceptions of the personas, or otherwise, one risks futile efforts with immature technology. In a sense, therefore, the question of picture quality for DDPs is also a question of feasibility study (of implementation).

Note that pictures constitute an essential element of the persona profile (Baxter, Courage, and Caine 2015; Nielsen et al. 2015). Pictures are instrumental for the persona to appear believable, and they have been found impactful for central persona perceptions, such as empathy (Probster, Haque, and Marsden 2018). Therefore, DDPs require these pictures in order to realise the many benefits associated with the use of personas in the HCI literature (Long 2009; Nielsen and Storgaard Hansen 2014).

To evaluate the *quality*, we first generate a sample of 1,000 artificial facial pictures using a state-of-the-art generator. To evaluate this sample, we then obtain 6,812 judgments from crowdworkers. To evaluate *user perceptions*, we conduct a 2×2 experiment with DDPs with a real/artificial picture. For measurement of user perceptions, we deploy the Persona Perception Scale (PPS) instrument (Salminen et al. 2018c) to gauge the impact of artificial pictures on the DDPs' *authenticity* and *clarity*, as well as the sense of *empathy*, and *willingness to use* among the online pool of respondents.

Thus, our research goal is to evaluate artificially generated pictures across multiple dimensions for deployment in DDPs. Note that our goal is *not* to make a technical AIG contribution. Rather, we apply a pre-existing method for persona profiles and then evaluate the results for user perceptions. So, our contribution is in the area of practical design and implementation of AIG.

Note also that even though we focus on DDPs in this research, many other domains and use cases have similar needs in terms of requiring large collections of diverse facial images, including HCI and human-robot interaction such as avatars (Ablanedo et al. 2018; Şengün 2014; Sengün 2015), robots (dos Santos et al. 2014; Duffy 2003; Edwards et al. 2016; Holz, Dragone, and O'Hare 2009), and chatbots (Araujo 2018; Go and Shyam Sundar 2019; Shmueli-Scheuer et al. 2018; Zhou et al. 2019a). Thus, our evaluation study has a cross-sectional value for other design purposes where artificial facial pictures would be useful.

2. Related literature

2.1. Lack of evaluation studies for artificial pictures

To quantify the need for evaluation studies of AIG in real systems, we carried out a scoping review (Bazzano

Figure 1. Data-driven persona development approach. Six-step process common for most DDP methods.

et al. 2017) by extracting information from 20 research articles that generate artificial facial pictures. The articles were retrieved via Google Scholar using relevant search phrases ('automatic image generation + faces', 'facial image creation', 'artificial picture generation + face', etc.) and focusing on peer-reviewed conference/ journal articles published between 2015 and 2019. The list of articles, along with the extracted evaluation methods, is provided in *Supplementary Material*.

Results show that evaluation methods in these articles almost always contain one or more technical metrics (90%, $N = 18$) and always a short, subjective evaluation by the authors (100%, $N = 20$), in the line of 'manual inspection revealed some errors but generally good quality' (not an actual quote). Among the 20 articles, less than half (45%, $N = 9$) measured actual human perceptions (typically using crowdsourced ratings). More importantly, none of the articles provided an evaluation study that would implement the generated pictures into a real system or application. The results of this scoping review thus show a general lack of user studies for practical evaluation of AIG in real systems or use cases (0% of the research we could locate did so).

As stated, the evaluation of AIG focuses on technical metrics (Gao et al. 2020) of image generation (e.g. inception score (Dey et al. 2019; Di and Patel 2017; Salimans et al. 2016; Yin et al. 2017), FID (Dey et al. 2019; Karras, Laine, and Aila 2019; Lin et al. 2019), Euclidean distance (Gecer et al. 2018), cosine similarity (Dey et al. 2019), reconstruction errors (Chen et al. 2019; Lee et al. 2018), or accuracy of face recognition (Di, Sindagi, and Patel 2018; Liu et al. 2017)). Many of the technical metrics are said to have various strengths and weaknesses (Barratt and Sharma 2018; Karras, Laine, and Aila 2019; Shmelkov, Schmid, and Alahari 2018; Zhang et al. 2018). The main weakness is that they do not capture user perceptions or UX ramifications of the pictures in real applications. This is because the technical metrics are not directly related to end-user experience when the user is observing the pictures within the *context* of their intended use (e.g. as part of DDPs).

Human evaluation studies, on the other hand, tend to focus on comparing the outputs of different algorithms, again ignoring the importance of context on the evaluation results. Typically, participants are asked to rank pictures produced using different algorithms from best to worst (Li et al. 2018; Liu et al. 2017; Zhou et al.

2019b) or rate the pictures by user perception metrics, such as realism, overall quality, and identity (Yin et al. 2017; Zhou et al. 2019b). For example, Li et al. (2018) recruited 84 volunteers to rank three generated images out of 10 non-makeup and 20 makeup test images based on quality, realism, and makeup style similarity. Lee et al. (2018) employed a similar approach by asking users which image is more realistic out of samples created using different generation methods. Similarly, Choi et al. (2018) asked crowd workers in Amazon Mechanical Turk (AMT) to rank the generated images based on realism, quality of attribute transfer (hair colour, gender, or age), and preservation of the person's original identity. The participants were shown four images at a time, generated using different methods. Zhang et al. (2018) conducted a two-alternative forced choice (2AFC) test by asking AMT participants which of the provided pictures is more similar to a reference picture.

On rarer instances, user perception metrics, such as realism, overall quality, and identity have been deployed. For example, Zhou et al. (2019) evaluated the quality of their generated results by asking if the participants consider the generated faces as realistic ('yes' or 'no'). In their study, 88.4% of the pictures were considered realistic. Iizuka, Simo-Serra, and Ishikawa (2017) recruited ten volunteers to evaluate the 'naturalness' of the generated pictures; the volunteers were asked to guess if a picture was real or generated. Overall, 77% of the generated pictures were deemed to be real. Yin et al. (2017) asked students to compare 100 generated pictures with original pictures along with three criteria: (1) *saliency* (the degree of the attributes that has been changed in the picture), (2) *quality* (the overall quality of the picture), and (3) *identity* (if the generated and the original picture are the same person). Their AIG method achieved an average quality rating of 4.20 out of 5. While these studies are closer to the realm of UX, we could not locate previous research that would (a) *investigate the effect of artificial pictures on UX of a real system*, or (b) *evaluate the impact of using artificially generated pictures on user perceptions*. However, evaluating AIG approaches for user perceptions and UX in real systems, is crucial for determining the success of AIG in real usage contexts for design, HCI, and various other areas of application (Özmen and Yucel 2019).
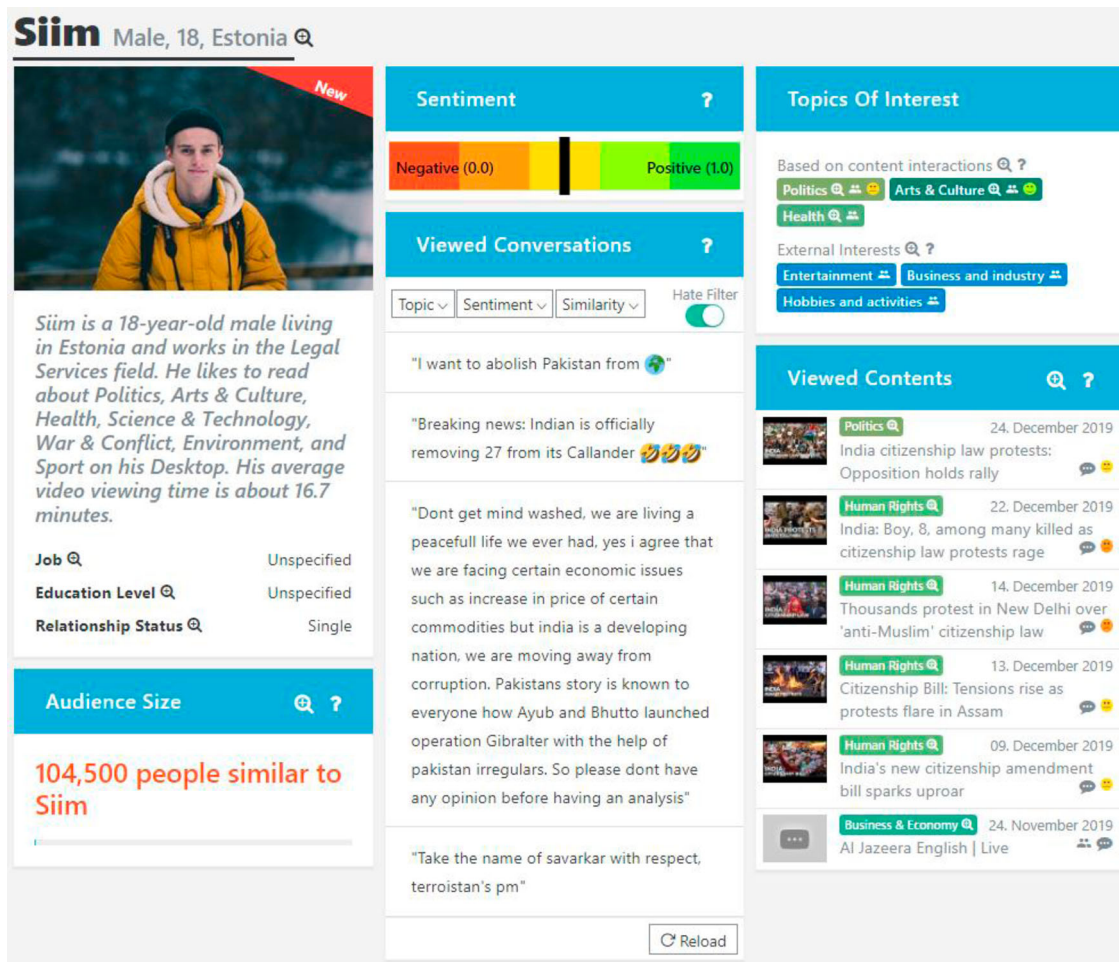
Figure 2. Example of DDP. The persona has a picture (stock photo in this example), name, age, text description, topics of interest, quotes, most viewed contents, and audience size. The picture is purchased and downloaded manually from an online photobank; the practical goal of this research is to replace manual photo curation through automatic image generation.

2.2. Data-driven persona development

A persona is a fictive person that describes a user or customer segment (Cooper 1999). Originating from HCI, personas are used in various domains, such as user experience/design (Matthews, Judge, and Whittaker 2012), marketing (Jenkinson 1994), and online analytics (Salminen et al. 2018b) to increase the empathy by designers, software developers, marketers (etc.) toward the users or customers of a product (Dong, Kelkar, and Braun 2007). Personas make it possible for decision makers to see use cases 'through the eyes of the user' (Goodwin 2009) and facilitate communication between team members through shared mental models (Pruitt and Adlin 2006). Researchers are increasingly developing methodologies for DDPs (McGinn and Kotamraju 2008; Zhang, Brown, and Shankar 2016) and automatic persona generation (An et al. 2018a; An et al. 2018b), mainly due to the increase in the availability of online user data and to increase the robustness of personas given the alternative forms of user understanding

(Jansen, Salminen, and Jung 2020). DDPs typically leverage quantitative social media and online analytics data to create personas that represent users or customers of a specific channel[1] (Salminen et al. 2017). Regarding the development of DDPs, for the generated pictures to be useful for personas, they need to be 'taken for real', meaning that they do not hinder the user perceptions of the personas (e.g. not reduce the persona's authenticity).

2.3. Persona user perceptions

Evaluation of user perceptions has been noted as a major concern of personas. Scholars have observed that personas need justification, mainly for their accuracy and usefulness in real organisations and usage scenarios (Chapman and Milham 2006; Friess 2012; Matthews, Judge, and Whittaker 2012). Prior research typically examines persona user perceptions via case studies (Faily and Flechais 2011; Jansen, Van Mechelen,

and Slegers 2017; Nielsen and Storgaard Hansen 2014), ethnography (Friess 2012), usability standards (Long 2009), or using statistical evaluation (An et al. 2018b; Brickey, Walczak, and Burgess 2012; Zhang, Brown, and Shankar 2016). For example, Friess (2012) investigated the adoption of personas among designers. Long (2009) measured the effectiveness of using personas as a design tool, using Nielsen's usability heuristics. Nielsen et al. (2017) analyze the match between journalists' preconceptions and personas created from the audience data, whereas Chapman et al. (2008) evaluate personas as quantitative information. While these evaluation approaches are interesting, survey methods provide a lucrative alternative for understanding how end users perceive personas. Survey research typically measures perceptions as latent constructs, apt for measurement of attitudes and perceptions that cannot be directly observed (Barrett 2007). This approach seems intuitively compatible with personas, as researchers have reported several attitudinal perceptions concerning personas (Salminen et al. 2019c). These are captured in the PPS survey instrument (Salminen et al. 2018c; Salminen et al. 2019f; Salminen et al. 2019g) that includes eight constructs and twenty-eight items to measure user perceptions of personas. We deploy this instrument in this research, as it covers essential user perceptions in the persona context.

2.4. Hypotheses

Following prior persona research, we formulate the following hypotheses to test persona user perceptions.

- H01: Using artificial pictures does not decrease the authenticity of the persona. HCI research has shown that authenticity (or *credibility*, *believability*) is a crucial issue for persona acceptance in real organisations — if the personas come across as 'fake', decision makers are unlikely to adopt them for use (Chapman and Milham 2006; Matthews, Judge, and Whittaker 2012). This is especially relevant for our context because personas already are fictitious people describing real user groups (An et al. 2018b), so we need to ensure that enhancing these fictitious people with artificially generated pictures does not further risk the perception of realism.
- H02: Using artificial pictures does not decrease the clarity of the persona profile. For personas to be useful, they should not be abstract or misleading (Matthews, Judge, and Whittaker 2012). HCI researchers have found that personas with inconsistent information make end users of personas confused (Salminen et al. 2018d; Salminen et al. 2019b). Again, we

need to ensure that artificial pictures do not make persona profiles more 'messy' or unclear for the end users.
- H03: Using artificial pictures does not decrease empathy towards the persona. Empathy is considered, among HCI scholars, as a key advantage of personas compared to other forms of presenting user data (Cooper 1999; Nielsen 2019). The generated personas need to 'resonate' with end users to make a real impact. Therefore, to be successful, artificial pictures should not reduce the sense of empathy towards the persona.
- H04: Using artificial pictures does not decrease the willingness to use the persona. Willingness to use (WTU) is a crucial construct for the adoption of personas for practical decision making (Rönkkö 2005; Rönkkö et al. 2004). HCI research has shown that if persona users do not show a willingness to learn more about the persona for their task at hand, persona creation risks remaining a futile exercise (Rönkkö et al. 2004).

Overall, ranking high on these perceptions is considered positive (desirable) within the HCI literature. This leads to defining the 'good enough' quality of artificial pictures in the DDP context such that *a 'good enough' picture quality does not decrease (a) the authenticity (i.e. the persona is still considered as 'real' as with real photographs), (b) clarity of the persona profile, (c) the sense of empathy felt toward the persona, or (d) the willingness to learn more about the persona.* In other words, it is the design goal of replacing real photographs with artificial pictures in the context of personas, with the concept being transferrable to other domains.

3. Methodology

3.1. Overview of evaluation steps

Our evaluation of picture quality consists of two separate studies: (1) crowdsourced evaluation study of AIG quality, and (2) user study measuring the perceptions of an online panel concerning personas with artificially generated pictures. The latter study tests if DDPs are perceived differently when using artificial pictures, while addressing the hypotheses presented in the previous section.

3.2. Research context

Our research context is a DDP system: *Automatic Persona Generation* (APG[2]). As a DDP system, APG requires thousands of realistic facial pictures to produce

a wide range of believable persona profiles for client organisations (Pruitt and Adlin 2006) covering a wide range of ages and ethnicities. An overview of the typical DDP development process is presented in Figure 3.

A practical limitation of APG is the need for manually acquiring facial pictures for the persona profiles (Salminen et al. 2019a). Because the pictures for APG are acquired from online stock photo banks (e.g. iStock-Photo, 123rf.com, etc.), manual effort is required to curate a large number of pictures. A large number of pictures is needed because APG can generate thousands of personas for client organisations – for each persona, a unique facial picture is required. Organisations over a lengthy period can have dozens of unique personas. Using stock photo banks also involves a financial cost (ranging from $1 to $20 USD per picture), making picture curation both time-consuming and costly. Given the goal of fully automated persona generation (Salminen et al. 2019a), there is a practical need for automatic image generation.

Thus, we evaluate the automatically generated facial pictures for use in APG (Jung et al. 2018a; Jung et al. 2018b). APG generates personas from online analytics and social media data. Figure 2 shows an example of a persona generated using the system. The practical purpose of automatically generated images is to replace the manual curation of persona profile pictures, saving time and money. Note that the cost and effort are not unique problems of APG, but generalise to all similar images systems, as the pictures need to be provided for each new persona generated.

3.3. Deploying StyleGAN for persona pictures

For AIG, we utilise a pre-trained version of *StyleGAN* (Karras, Laine, and Aila 2019), a state-of-the-art generator that represents a leap towards photorealistic facial pictures and can be freely accessed on GitHub.[3] StyleGAN was chosen for this research because (a) it

is a leap toward generating photorealistic facial images, especially relative to the previous state-of-art, (b) the trained model is publicly available, and (c) its deployment is robust for possible use in real systems. Style-GAN generated the images, so this is a back end process.

We use a pretrained model from the creators of StyleGAN (Karras, Laine, and Aila 2019). This model was trained on CelebA-HQ and FFHQ datasets using eight Tesla V100 GPUs. It is implemented in TensorFlow,[4] an open-source machine learning library and is available in a GitHub repository.[5] We access this pre-trained model via the GitHub repository that contains the model and the required source code to run it.

Our goal is to use this pre-trained model to generate a sample of 1,000 realistic facial pictures. The method of applying the published code to generate the pictures is straightforward. We provide the exact steps below to facilitate replication studies:

- **Step 1:** Import the required Python packages (*os*, *pickle*, *numpy*, from *PIL*: *Image*, *dnnlib*).

- **Step 2:** Define the parameters and paths

- **Step 3:** Initialize the environment and load the pre-trained StyleGAN model.

- **Step 4:** Set random states and generate new random input. Randomization is needed because the model always generates the same face for a particular input vector. To generate unique images, a unique set of input arrays should be provided. This is done by setting a random state equal to the current number of iterations, which allows us to have unique images and reproducible results at the same time.

- **Step 5:** Generate images using the random input array created in the previous step.

- **Steps 6:** Save the generated images as files to the output folder. We use the resolution of 1024 × 1024 pixels. Other available resolutions are 512 × 512 px and 256 × 256 px.
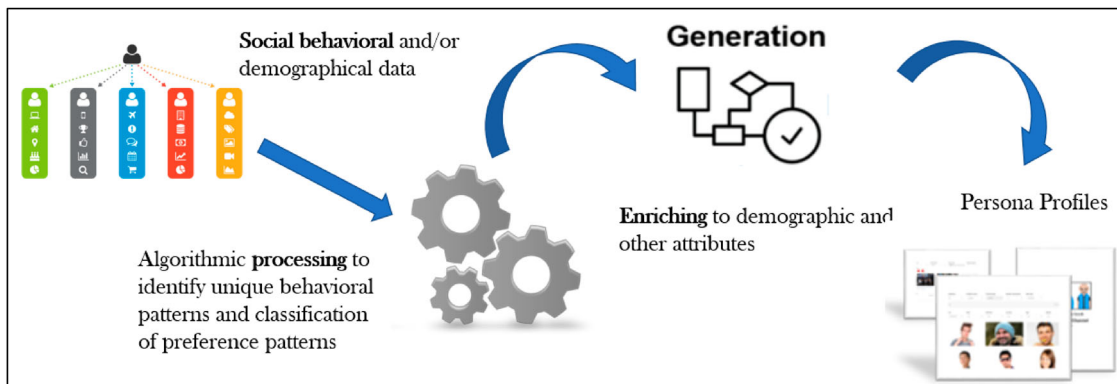

Figure 3. APG data and processing flowchart from server configuration to data collection and persona generation.

The above steps with the mentioned parameters enable us to generate artificial pictures with similar quality to those in the StyleGAN research paper (Karras, Laine, and Aila 2019). For replicability, we are sharing the Python code we used for AIG in *Supplementary Material*.

4. STUDY 1: crowdsourced evaluation

4.1. Method

We evaluate the human-perceived quality of 1,000 generated facial pictures. To facilitate comparison with prior work using human evaluation for artificial pictures (Choi et al. 2018; Song et al. 2019; Zhang et al. 2018), we opt for crowdsourcing, using Figure Eight to collect the ratings. This platform has been widely used for gathering manually annotated training data (Alam, Ofli, and Imran 2018) and ratings (Salminen et al. 2018a) in various subdomains of computer science. The pictures were shown in the full 1024×1024 pixels format to provide the crowd raters enough detail for a valid evaluation. The following task description was provided to the crowd raters, including the quality criteria and examples:

You are shown a facial picture of a person. Look at the picture and choose how well it represents a real person. The options:

- 5: Perfect—the picture is indistinguishable from a real person.
- 4: High quality—the picture has minor defects, but overall it's pretty close to a real person.
- 3: Medium quality—the picture has some flaws that suggest it's not a real person.
- 2: Low quality—the picture has severe malformations or defects that instantly show it's a fake picture.
- 1: Unusable—the picture does not represent a person at all.

We also clarified to the participants that the use case is to find realistic pictures specifically for persona profiles, explaining that these are descriptive people of some user segment. Additionally, we indicated in the title that the task is to evaluate *artificial* pictures of people, to manage the expectations of the crowd raters accordingly (Pitkänen and Salminen 2013). Other than the persona aspect, these are similar to guidelines used in prior work to facilitate image comparisons.

Following the quality control guidelines for crowdsourcing by Huang, Weber, and Vieweg (2014) and Alonso (2015), we implemented suitable parameters in the Figure Eight platform. We also enabled *Dynamic judgments*, meaning the platform automatically collects more ratings when there is a higher disagreement among the raters. Based on the results of a pilot study with 100 pictures, not used in the final research, we set the maximum number of ratings to 5 and confidence goal to 0.65. The default number of raters was three, so the platform only went to 5 raters if a 0.65 confidence was not achieved.[6]

4.2. Results

We spent $266.98 USD to obtain 6,812 crowdsourced image ratings. This was the number of evaluations from trusted contributors, not including the test questions. Note that if the accuracy of a crowd rater's ratings relative to the test questions falls below the minimum accuracy threshold (in our case, 80%), the rater is disqualified, and the evaluations become untrusted. There were 423 untrusted judgments (6% of the total submitted ratings), i.e. ratings coming from contributors that continuously fail to correctly rate the test pictures. Thus, 94% of the total ratings were deemed trustworthy. The majority label for each rated picture is assigned by comparing the confidence-adjusted ratings of each available class, calculated as follows:

$$\text{Confidence}_{\text{class}} = \frac{\sum_{i=1}^{n} \text{trust}_{\text{class}}}{\sum_{i=1}^{n} \text{trust}_{\text{all}}},$$
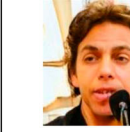
where the confidence score of the class is given by the sum of the trust scores from all n raters of that picture. The trust score is based on a crowdworker's historical accuracy (relative to test questions) on all the jobs he/she has participated in. For example, if the confidence score of 'perfect' is 0.66 and 'medium quality' is 0.72, then the chosen majority label is 'medium quality' (0.72 > 0.66).

The results (see Table 1) show 'High quality' as the most frequent class. Sixty percent (60%) of the generated pictures are rated as either 'Perfect' or 'High quality'. The average quality score was 3.7 out of 5 (SD = 0.91) when calculated from majority votes and 3.8 when calculated from all the ratings. 9.9% of the pictures were rated as 'Low quality', and none was rated as 'Unusable'.

4.3. Reliability analysis

To assess the reliability of the crowd ratings, we measured the interrater agreement of the quality ratings among crowdworkers. For this, we used two metrics: *Gwet's AC1* (AC1) and *percentage agreement* (PA). Using AC1 is appropriate when the outcome is ordinal, the number of ratings varies across items (Gwet 2008) and where the Kappa metric is low despite a high level

Table 1. The results of crowd evaluation based on a majority vote of the picture quality. Most frequent class bolded. Example facial image from each of the 5 classes shown for comparison.

5	4	3	2	1	total
Perfect	*High quality*	*Medium quality*	*Low quality*	*Unusable*	
				n/a	-
N = 214	**N=388**	N = 298	N = 100	N = 0	1000
21.4%	**38.8%**	29.8%	10.0%	0.0%	100%

of agreement (Banerjee et al. 1999; Salminen et al. 2018a). Because of these properties, we chose AC1 with ordinal weights as the interrater agreement metric. In addition, PA was calculated as a simple baseline measure. Standard errors were used to construct the 95% confidence interval (CI) for AC1. For PA, 95% CI was calculated using 100 bootstrapped samples.

Results (see Table 2) show a high PA agreement (86.2%). The interrater reliability was 0.627, in the range of *good* (i.e. 0.6−0.8) (Wongpakaran et al. 2013). The results were statistically significant ($p < 0.001$), with the probability of observing such results by chance is less than 0.1%. Therefore, the crowd ratings can be considered to have satisfactory internal validity. However, the quality of some pictures is more easily agreed upon than others. When stratified, the overall agreement and AC1 were similar across *low, moderate,* and *high* quality labels (PA ∼ 85%, and AC1 ∼ 0.75). However, the agreement was lower when the picture was rated perfect (PA = 76.7%, AC1 = 0.498). This implies that 'perfect' is more difficult to determine than the other rating labels.

5. STUDY 2: effects on persona perceptions

5.1. Experiment design

We created two base personas using the APG (An et al. 2018b) methodology described previously; one male and one female. We leave the evaluation of other genders for future research. The experiment variable is the use of an artificial image in the persona profile.

Table 2. Agreement metrics for the crowdsourced ratings showing satisfactory internal validity.

Measure	Value	SE	95% CI	P
PA	86.2%	0.6%	85.27%, 87.2%	
AC1	0.627	0.017	0.59, 0.66	<0.001

The other elements of the persona profiles are identical between the two treatments. For this, we manipulated the base personas by introducing either (a) *a real photograph of a person* or (b) a *demographically matching artificial picture* (see Figure 4).

The demographic match was determined manually by two researchers who judged that the chosen pictures were similar for gender, age, and race. Using a modified Delphi method, a seed image of either a real or article picture was select using the meta-data attributes of gender, age, and race. The researchers independently selected matching images for each. The two researchers then jointed selected the mutually agreed upon image for the treatments. The artificial pictures were chosen from the ones rated 'perfect' by the crowd raters. The real photos were sourced from online repositories, with Creative Commons license.

In total, four persona treatments were created: *Male Persona with Real Picture* (MPR), *Male Persona with Artificial Picture* (MPA), *Female Persona with Real Picture* (FPR), and *Female Persona with Artificial Picture* (FPA). The created personas were mixed into four sequences:

- **Sequence 1:** MPR → FPA
- **Sequence 2:** MPA → FPR
- **Sequence 3:** FPR → MPA
- **Sequence 4:** FPA → MPR

Each participant was randomly assigned to one of the sequences. To counterbalance the dataset, we ensured an even number of participants ($N = 520/4 = 130$) for each sequence. Technically, the participants self-selected the sequence, as each participant could only take one survey. The participants were excluded from answering in more than one survey based on their (anonymous) Respondent ID. The gender distribution for each of the four sequences, as shown: S1 (*M:*
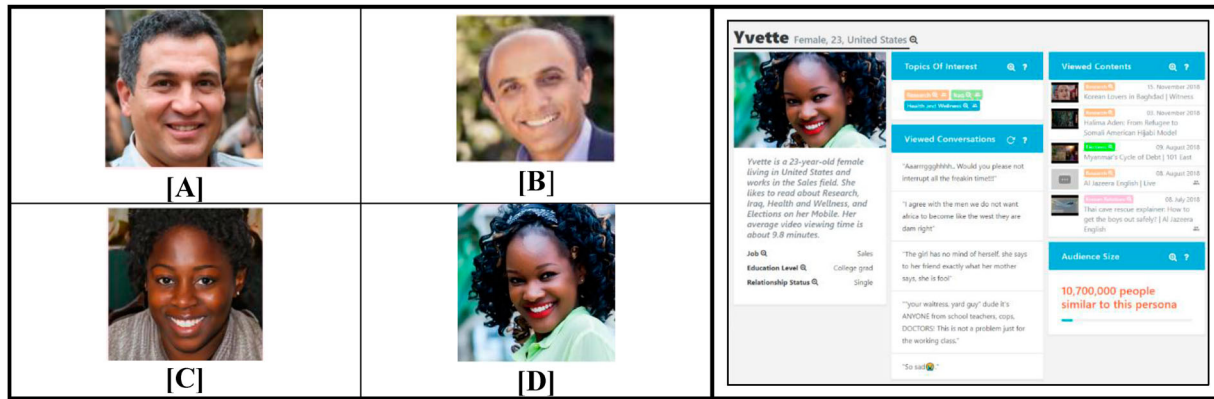
Figure 4. Artificial male picture [A], Real male picture [B], Artificial female picture [C], and Real female picture [D]. Among the male/female personas, all other content in the persona profile was the same except the picture that alternated between Artificial and Real. Pictures of the full persona profiles are provided in Supplementary Material.

41.5% *F*: 58.5%), S2 (*M*: 39.0% *F*: 61.0%), S3 (*M*: 39.8% *F*:60.2%), S4 (*M*: 34.5% *F*: 64.5%).

5.2. Recruitment of participants

We created a survey for each sequence. In each survey, we (a) explain to participants what the research is about and what personas are ('*a persona is defined as a fictive person describing a specific customer group*'). Then, we (b) show an example persona with explanations of the content, and (c) explain the task scenario ('*Imagine that you are creating a YouTube video for the target group that the persona you will be shown next describes*'.). After this, (d) the participants are shown one of the four treatments, asked to review the information carefully, and complete the PPS questionnaire.

In total, 520 participants were recruited using Prolific, an online survey platform often applied in social science research (Palan and Schitter 2018). Prolific was chosen for this evaluation step (as opposed to previously used Figure Eight), as it provides background information of the participants (e.g. gender, age) that can be deployed for further analyses. The average age of the participants was 35 years old (SD = 7.2), with 59.1% being female, overall. The nationality of the participants was the United Kingdom, they had at least an undergraduate degree, and none were students. We verified the quality of the answers using an attention check question ('*It's important that you pay attention to this study. Please select "Slightly agree"'*.). Out of 520 answers, 19 (3.7%) failed the attention check; these answers were removed. In addition, five answers were timed out by the Prolific platform. Therefore, we ended up with 496 qualified participants (95.4% of the total).

5.3. Measurement

The perceptions are measured using the PPS (Salminen et al. 2018c), a survey instrument measuring what individuals think about specific personas (see Table 3). The PPS has previously deployed in several persona experiments (see [Salminen et al. 2019f; Salminen et al. 2019g]). Note that the *authenticity* construct is similar to constructs in earlier artificial image evaluation – specifically to *realism* (Zhou et al. 2019b) and *naturalness* (Iizuka, Simo-Serra, and Ishikawa 2017). However, the other constructs expand the perceptions typically used for image evaluation. In this sense, the hypotheses (a) add novelty to the measurement of user perceptions regarding the employment of artificial images in a real system output, and (b) are relevant for the design and use of personas.

5.4. Analysis procedure

The participants were grouped based on the persona presented (either the male or the female one), and whether the persona picture was artificial or real. The data was re-arranged to disentangle the gender of the persona, leading to one male-persona dataset (with a 'real' and an 'artificial' group), and a female-persona dataset with similar groups. This allowed the usage of a standard MANOVA (Hair et al. 2009) to determine whether the measurements differed across artificial and real pictures; both genders were analysed independently.

To enhance the robustness of the findings, Bayesian independent samples tests were used to estimate Bayesian Factors (BF), comparing the likelihoods between the null and alternative hypotheses (Lee 2014). A Naïve Bayes approach was employed with regards to priors.

Table 3. Survey statements. The participants answered using a 7-point Likert scale, ranging from Strongly Disagree to Strongly Agree. The statements were validated in (Salminen et al. 2018c). WTU – willingness to use.

Perception	Statements
Authenticity	The persona seems like a real person. I have met people like this persona. The picture of the persona looks authentic. The persona seems to have a personality.
Clarity	The information about the persona is well presented. The text in the persona profile is clear enough to read. The information in the persona profile is easy to understand.
Empathy	I feel like I understand this persona. I feel strong ties to this persona. I can imagine a day in the life of this persona.
Willingness To Use	I would like to know more about this persona. This persona would improve my ability to make decisions about the customers it describes. I would make use of this persona in my task [of creating a YouTube video].

5.5. Results

Male persona: Beginning with the multivariate tests, no significant effects were registered for Type of Picture (Pillai's Trace = 0.017, $F(5, 476) = 1.651$, $\eta_p^2 = 0.017$, $p = 0.145$), indicating that none of the measurements differed across male real and artificial pictures for the persona profile. Nevertheless, we proceeded with an analysis of univariate tests, which confirmed that none of the measurements differed across types of pictures. The univariate differences for between-subjects are summarised in Table 4.

The lack of differences in scale ratings (see Figure 5) also indicates that the use of real or artificial pictures results in no differences for *authenticity*, *clarity*, *empathy*, or *willingness to use* for the male persona.

The Bayesian analysis on the male persona indicates strong lack of evidence for differences regarding *clarity* (BF = 13.856; $F(1, 480) = 0.002$; $p = 0.965$) and *willingness to use* (BF = 10.030; $F(1, 480) = 0.658$; $p = 0.418$), and moderate lack of evidence for differences regarding *authenticity* (BF = 5.126; $F(1, 480) = 2.023$; $p = 0.156$) and *empathy* (BF = 5.040; $F(1, 480) = 2.057$; $p = 0.152$) (Jeffreys 1998).

Female persona: Beginning with the multivariate tests, unlike with the male persona, significant effects were registered for Type of Picture (Pillai's Trace = 0.051, $F(5, 476) = 5.081$, $\eta_p^2 = 0.051$, $p < 0.001$),

indicating that at least one of the measurements differed between real and artificial pictures for the female persona. Thus, we proceeded with univariate testing to determine which of the measurements exhibited differences across picture type (see Table 4). *Authenticity* had significant differences across types of picture (BF = 0.032; $F(1, 480) = 12.479$, $p < 0.001$). Artificial female pictures were perceived as more *authentic* ($M = 5.075$, SD = 1.016) than real pictures ($M = 4.711$, SD = 1.235). None of the other measurements differed across types of pictures. Figure 6 illustrates the comparison between the two groups for the female persona.

This was corroborated by the Bayesian Factors that indicate that strong lack of evidence regarding differences for *clarity* (BF = 13.865; $F(1, 480) = 0.001$, $p = 0.980$) and *willingness to use* (BF = 13.828; $F(1, 480) = 0.006$, $p = 0.938$), and moderate lack of evidence for *empathy* (BF = 8.290; $F(1, 480) = 1.045$, $p = 0.307$) (Jeffreys 1998).

Finally, as one the statements in PPS specifically dealt with the picture of the persona (Item 3: '*The picture of the persona looks authentic*'.), we inspected the mean scores of this statement separately. In line with our other findings, the artificial female persona picture is in fact considered to be more authentic than the real photograph ($M_{FPA} = 5.74$ vs. $M_{FPR} = 4.89$). This difference is statistically significant ($t(480) = 6.896$, $p < 0.001$). For

Table 4. Univariate tests for between-subjects effects (df(error) = 1(480)).

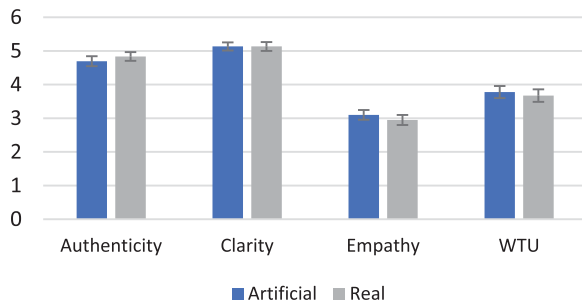| Independent variable	Dependent variable	*Male* persona			*Female* persona		
		F	η_p^2	*p*-value	F	η_p^2	*p*-value
Type of Picture (real or artificial)	Authenticity	2.023	0.004	0.156	12.479	0.025	<0.001
	Clarity	0.002	<0.001	0.965	0.001	<0.001	0.980
	Empathy	2.057	0.004	0.152	1.045	0.002	0.307
	WTU	0.658	0.001	0.418	0.006	<0.001	0.938

Figure 5. Means of the scale variables for the *male* persona. Error bars indicate standard error.

the male persona, differences are minimal ($M_{MPA} = 5.11$ vs. $M_{MPR} = 5.14$) and not statistically significant ($t(480) = -0.187$, $p = 0.851$).

In summary, for H01, there were no significant differences in the perceptions for the male persona; however, for the female persona, artificial pictures actually *increased* the perceived authenticity. For the other perceptions, there was no significant change when replacing the real photo with the artificially generated picture. Therefore,

- There was no evidence that using artificial pictures decrease the perceived authenticity of the persona (H01: supported).
- There was no evidence that using artificial pictures decrease the clarity of the persona profile (H02: supported).
- There was no evidence that using artificial pictures decrease empathy towards the persona (H03: supported).
- There was no evidence that using artificial pictures decrease the willingness to use the persona (H04).

6. Discussion

6.1. Can artificial pictures be used for DDPs?

Our analysis focuses on a timely problem in a relevant, yet underexplored area. However, it is one of increasing
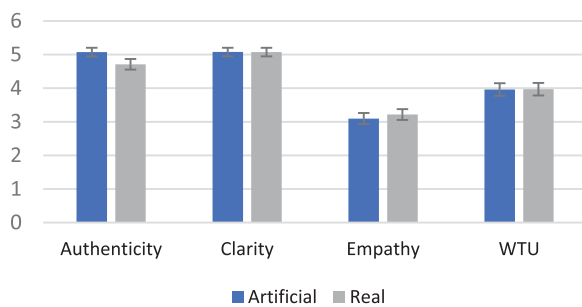

Figure 6. Means of the scale variables for the *female* persona. Error bars indicate standard error.

importance in a media rich online environment (Church, Iyer, and Zhao 2019). The impact of artificial facial pictures on user perceptions has not been studied thoroughly in previous HCI design literature. The lack of applied user studies is understandable given that until recently, the generated facial pictures were not close to realistic, so the research focus was on improving algorithms. However, as the quality of the facial pictures improves, the focus ought to shift towards evaluation studies in real-world use cases, systems, and applications. As there is a lack of literature in this regard, the research presented here contains a step forward in analysing the use of artificially facial generated pictures in real systems.

In terms of results, the crowd evaluation suggests that more than half of the artificial pictures are considered as either *perfect* or *high quality*. The ratio of 'perfect and high-quality' pictures to the rest is around 1.5, implying that most of the pictures are satisfactory according to the guidelines we provided. The persona perception analysis shows that the use of artificial pictures vs. real pictures in persona profiles does not reduce the authenticity of the persona or people's willingness to use the persona, two crucial concerns of persona applicability. Therefore, we find the state-of-the-art of AIG satisfactory for a persona and most likely for other systems requiring the substantial use of facial images. So, it is possible to replace the need for manually retrieving pictures from online photo banks with a process of automatically generated pictures.

6.2. Gender differences in perception

Regarding the female persona with an artificial picture being perceived as more authentic, we surmise that there might be a 'stock photo' effect involved, rather than a gender effect. This proposition is backed up by previous findings of stock photos being perceived differently by individuals than non-stock photos (Salminen et al. 2019f). Visually, to the respondents, the real photo chosen for the female persona appears different from the one chosen for the male persona (see Figure 4). It is difficult to explain or quantify why this is. We interpret this finding such that the choice of pictures for a persona profile, and perhaps other system contexts, is a delicate matter; even small nuances can affect user perceptions.

This interpretation is generally in line with previous HCI research regarding the foundational impact of photos in persona profiles (Hill et al. 2017; Salminen et al. 2018d; Salminen et al. 2019b). Possibly, stock photos can appear, at times, less realistic than photos of 'real people' because they are 'too shiny, too perfect'

(or 'too smiling' [Salminen et al. 2019e]). Thus, *if the generator's outputs are closer to real people than stock photos in their appearance*, it is possible that these pictures are deemed more realistic than stock photos. However, this does not explain why the effect was found for the female persona and not for the male one. The only way to establish if there is a gender effect that influences perceptions of stock photos is to conduct repeated experiments with stock photos of different people. In addition to repeated experiments, for future research, the gender difference suggests another variable to consider: 'the degree of photo-editing' or 'shiny factor' (i.e. how polished the stock photo is and how this affects persona perceptions). The proper adjustment of this variable is best ensured via a manipulation check.

6.3. Wider implications for HCI and system implementation

In a broader context, the manuscript contributes to evaluating a machine learning tool for UX/UI design work. For the use of artificial images, guidelines are provided.

- Solutions for Mitigating Subjectivity: The results indicate that evaluating the quality of artificial facial pictures contains a moderate to a high level of subjectivity, making reliable evaluation for production systems costlier. We hypothesise that there will always be some degree of subjectivity, as individuals vary in their ability to pay attention to details. This can be partially remedied by *choosing the pictures with the highest agreement between the raters*, or *using a binary rating scale* (i.e. 'good enough' vs. 'not good enough') as the agreement is generally easier to obtain with fewer classes (Alonso 2015). The observed 'disagreement' may be partly fallacious because people might agree whether a picture is either usable (4 or 5) or non-usable (1 or 2), but the exact agreement between 4 or 5, for example, is lower. As stated, for practical purposes, it does not matter if a picture is 'Perfect' or 'High quality', as both classes are decent, at least for this use case.
- Handling of Borderline Cases. Regarding pictures to use in a production system, we recommend a *borderline principle*: if in doubt of the picture quality, reject it. The marginal cost of generating new pictures is diminishingly low but showing a low-quality picture decreases user experience, sometimes drastically. For this reason, the economics of automatic image generation are in favour of rejecting borderline images more than letting through distorted images. However, rejecting borderline images does increase the

total cost of evaluation because to obtain n useful pictures, one now has to obtain $n \times (1 + \text{false positive rate})$ ratings, which is $(n \times (1 + \text{false positive rate}) - n) / n$ ratings more than n ratings. Additionally, as we have shown, the higher the disagreement among the crowd raters, the more ratings required.
- Final Choice for Human. In evaluating the suitability of artificial pictures for use in real applications, domain expertise is needed because, irrespective of quality guidelines, the crowd may have different quality standards than domain experts. For example, the crowd can be used to filter out low-quality photos, but the 'better' quality photos should be evaluated specifically by domain experts, as different domains likely have different quality standards. For personas, the pictures need to be of high quality, but when implementing them for the system, they are cropped into a smaller resolution that helps obfuscate minor errors.

6.4. Future research avenues

The following avenues for future research are proposed.

- Suitability in Other Domains. For example, *how do quality standards and requirements by users and organizations differ across domains and use cases? How well are artificial ('fake') pictures detected by end users, such as consumers and voters?* This research ties in with the nascent field of 'deep fakes' (Yang, Li, and Lyu 2019), i.e. images and videos purposefully manipulated for a political or commercial agenda. To this end, future studies could investigate the wider impact of using AI-generated images for profile pictures on sharing, economy platforms, or social media and news sites, and how that impact user perceptions, such as trust. Another interesting domain for suitability studies includes marketing, as facial pictures are widely deployed to advertise products such as fashion and luxury items.
- Algorithmic Bias. It would be important to investigate if the generated pictures involve an algorithmic bias – given that the training data may be biased, it would be worthwhile to analyze how diverse the generated pictures for different ethnicities, ages, and genders. Regarding persona perceptions, the race could be a confounding factor in our research and should be analysed separately in future research. A related question is: does the picture quality vary by demographic factors such as gender and race? Studies on algorithmic bias have been carried out within the HCI community (Eslami et al. 2018; Salminen et al. 2019b) and should be extended to this context.

- Demographically Conditional Images. For future development, we envision a system that automatically generates persona-specific pictures based on specific features/attributes of the personas – this would enable 'on-demand' picture creation for new personas generated by APG, whereas currently, the pictures need to manually tagged for age, gender, and country.

7. Conclusion

Our research goal was to evaluate the applicability of artificial pictures for personas along two dimensions: their quality and their impact on user perceptions. We found that more than half of the pictures were rated as perfect or high quality, with none as unusable. Moreover, the use of artificial pictures did not decrease the perceptions of personas that are found important in the HCI literature. These results can be considered as a vote of confidence for the current state of technology concerning the automatic generation of facial pictures and their use in data-driven persona profiles.

Notes

1. A demo of the system can be accessed at https://persona.qcri.org
2. https://persona.qcri.org
3. https://github.com/NVlabs/stylegan
4. https://www.tensorflow.org/
5. https://github.com/NVlabs/stylegan
6. Confidence is defined as agreement adjusted by trust score of each rater.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Ablanedo, J., E. Fairchild, T. Griffith, and C. Rodeheffer. 2018. "Is This Person Real? Avatar Stylization and Its Influence on Human Perception in a Counseling Training Environment." In: Chen J., Fragomeni G. (eds). In *International Conference on Virtual, Augmented and Mixed Reality*, 279–289. Lecture Notes in Computer Science, vol 10909. Springer, Cham. doi:10.1007/978-3-319-91581-4_20.

Alam, F., F. Ofli, and M. Imran. 2018. "Crisismmd: Multimodal Twitter Datasets from Natural Disasters." In *Twelfth International AAAI Conference on Web and Social Media*. AAAI. Palo Alto, California, USA.

Alonso, O. 2015. "Practical Lessons for Gathering Quality Labels at Scale." In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1089–1092. doi:10.1145/2766462.2776778.

An, J., H. Kwak, S. Jung, J. Salminen, and B. J. Jansen. 2018a. "Customer Segmentation Using Online Platforms: Isolating Behavioral and Demographic Segments for Persona Creation via Aggregated User Data." *Social Network Analysis and Mining* 8 (1). doi:10.1007/s13278-018-0531-0.

An, J., H. Kwak, J. Salminen, S. Jung, and B. J. Jansen. 2018b. "Imaginary People Representing Real Numbers: Generating Personas From Online Social Media Data." *ACM Transactions on the Web (TWEB)* 12 (4): Article No. 27. doi:10.1145/3265986.

Antipov, G., M. Baccouche, and J.-L. Dugelay. 2017. "Face Aging with Conditional Generative Adversarial Networks." In *2017 IEEE International Conference on Image Processing (ICIP)*, IEEE, Beijing, 2089–2093.

Araujo, T. 2018. "Living up to the Chatbot Hype: The Influence of Anthropomorphic Design Cues and Communicative Agency Framing on Conversational Agent and Company Perceptions." *Computers in Human Behavior* 85: 183–189. doi:10.1016/j.chb.2018.03.051.

Ashraf, M., N. I. Jaafar, and A. Sulaiman. 2019. "System- vs. Consumer-Generated Recommendations: Affective and Social-Psychological Effects on Purchase Intention." *Behaviour & Information Technology* 38 (12): 1259–1272. doi:10.1080/0144929X.2019.1583285.

Banerjee, M., M. Capozzoli, L. McSweeney, and D. Sinha. 1999. "Beyond Kappa: A Review of Interrater Agreement Measures." *Canadian Journal of Statistics* 27 (1): 3–23. doi:10.2307/3315487.

Barratt, S., and R. Sharma. 2018. A Note on the Inception Score. *ArXiv:1801.01973 [Cs, Stat]*. http://arxiv.org/abs/1801.01973.

Barrett, P. 2007. "Structural Equation Modelling: Adjudging Model fit." *Personality and Individual Differences* 42 (5): 815–824.

Baxter, K., C. Courage, and K. Caine. 2015. *Understanding Your Users: A Practical Guide to User Requirements Methods, Tools, and Techniques*. 2nd ed. Burlington, Massachusetts. Morgan Kaufmann.

Bazzano, A. N., J. Martin, E. Hicks, M. Faughnan, and L. Murphy. 2017. "Human-centred Design in Global Health: A Scoping Review of Applications and Contexts." *PloS One* 12 (11).

Brangier, E., and C. Bornet. 2011. "Persona: A Method to Produce Representations Focused on Consumers' Needs." Eds: Waldemar Karwowski, Marcelo M. Soares, Neville A. Stanton. In *Human Factors and Ergonomics in Consumer Product Design*, 37–61. Taylor and Francis.

Brauner, P., R. Philipsen, A. C. Valdez, and M. Ziefle. 2019. "What Happens when Decision Support Systems Fail? — The Importance of Usability on Performance in Erroneous Systems." *Behaviour & Information Technology* 38 (12): 1225–1242. doi:10.1080/0144929X.2019.1581258.

Brickey, J., S. Walczak, and T. Burgess. 2012. "Comparing Semi-Automated Clustering Methods for Persona Development." *IEEE Transactions on Software Engineering* 38 (3): 537–546.

Chapman, C. N., E. Love, R. P. Milham, P. ElRif, and J. L. Alford. 2008. "Quantitative Evaluation of Personas as Information." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 52 (16): 1107–1111. doi:10.1177/154193120805201602.

Chapman, C. N., and R. P. Milham. 2006. "The Personas' New Clothes: Methodological and Practical Arguments against a Popular Method." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50 (5): 634–636. doi:10.1177/154193120605000503.

Chen, A., Z. Chen, G. Zhang, K. Mitchell, and J. Yu. 2019. "Photo-Realistic Facial Details Synthesis from Single Image." In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 9428–9438. doi:10.1109/ICCV.2019.00952.

Choi, Y., M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. 2018. "Stargan: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, Salt Lake City, UT. 8789–8797.

Church, E. M., L. Iyer, and X. Zhao. 2019. "Pictures Tell a Story: Antecedents of Rich-Media Curation in Social Network Sites." *Behaviour & Information Technology* 38 (4): 361–374. doi:10.1080/0144929X.2018.1535620.

Cooper, A. 1999. *The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity*. 1st ed. Carmel, Indiana. Sams – Pearson Education.

Cooper, A. 2004. *The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity*. 2nd ed. Carmel, Indiana. Pearson Higher Education.

Dey, R., F. Juefei-Xu, V. N. Boddeti, and M. Savvides. 2019. "RankGAN: A Maximum Margin Ranking GAN for Generating Faces." In *Computer Vision – ACCV 2018. (Vol. 11363)*. Cham: Springer. doi:10.1007/978-3-030-20893-6_1.

Di, X., and V. M. Patel. 2017. "Face Synthesis from Visual Attributes via Sketch Using Conditional VAEs and GANs." *ArXiv:1801.00077 [Cs]*. http://arxiv.org/abs/1801.00077.

Di, X., V. A. Sindagi, and V. M. Patel. 2018. "GP-GAN: Gender Preserving GAN for Synthesizing Faces from Landmarks." In *2018 24th International Conference on Pattern Recognition (ICPR)*, 1079–1084. doi:10.1109/ICPR.2018.8545081.

Dong, J., K. Kelkar, and K. Braun. 2007. "Getting the Most Out of Personas for Product Usability Enhancements." In *Usability and Internationalization. HCI and Culture*, 291–296. http://www.springerlink.com/index/C0U2718G14HG1263.pdf.

dos Santos, T. F., D. G. de Castro, A. A. Masiero, and P. T. A. Junior. 2014. "Behavioral Persona for Human-Robot Interaction: A Study Based on Pet Robot Kurosu M. (eds) In. *International Conference on Human-Computer Interaction*, 687–696.

Duffy, B. R. 2003. "Anthropomorphism and the Social Robot." *Robotics and Autonomous Systems* 42 (3–4): 177–190.

Edwards, A., C. Edwards, P. R. Spence, C. Harris, and A. Gambino. 2016. "Robots in the Classroom: Differences in Students' Perceptions of Credibility and Learning Between 'Teacher as Robot' and 'Robot as Teacher'." *Computers in Human Behavior* 65: 627–634. doi:10.1016/j.chb.2016.06.005.

Eslami, M., S. R. Krishna Kumaran, C. Sandvig, and K. Karahalios. 2018. "Communicating Algorithmic Process in Online Behavioral Advertising." In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA, Paper 432.

Faily, S., and I. Flechais. 2011. "Persona Cases: A Technique for Grounding Personas." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2267–2270. doi:10.1145/1978942.1979274.

Friess, E. 2012. "Personas and Decision Making in the Design Process: An Ethnographic Case Study." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1209–1218. doi:10.1145/2207676.2208572.

Gao, F., J. Zhu, H. Jiang, Z. Niu, W. Han, and J. Yu. 2020. "Incremental Focal Loss GANs." *Information Processing & Management* 57 (3): 102192. doi:10.1016/j.ipm.2019.102192.

Gecer, B., B. Bhattarai, J. Kittler, and T.-K. Kim. 2018. "Semi-supervised Adversarial Learning to Generate Photorealistic Face Images of New Identities from 3D Morphable Model." In *Computer Vision – ECCV 2018. ECCV 2018 (Vol. 11215)*, 230–248. Cham: Springer. doi:10.1007/978-3-030-01252-6_14.

Go, E., and S. Shyam Sundar. 2019. "Humanizing Chatbots: The Effects of Visual, Identity and Conversational Cues on Humanness Perceptions." *Computers in Human Behavior*. doi:10.1016/j.chb.2019.01.020.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. 2014. "Generative Adversarial Nets." In *Advances in Neural Information Processing Systems 27*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, 2672–2680. Curran Associates, Inc. http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf.

Goodwin, K. 2009. *Designing for the Digital Age: How to Create Human-Centered Products and Services*. New York, New York. 1st ed. Wiley.

Gwet, K. L. 2008. "Computing Inter-Rater Reliability and Its Variance in the Presence of High Agreement." *British Journal of Mathematical and Statistical Psychology* 61 (1): 29–48. doi:10.1348/000711006X126600.

Hair, J. F., W. C. Black, B. J. Babin, and R. E. Anderson. 2009. *Multivariate Data Analysis*. 7th ed. New York, New York. Pearson.

Heusel, M., H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. 2017. "Gans Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium." *Advances in Neural Information Processing Systems*, 6626–6637.

Hill, C. G., M. Haag, A. Oleson, C. Mendez, N. Marsden, A. Sarma, and M. Burnett. 2017. "Gender-Inclusiveness Personas vs. Stereotyping: Can We Have It Both Ways?" In *Proceedings of the 2017 CHI Conference*, 6658–6671. doi:10.1145/3025453.3025609.

Holz, T., M. Dragone, and G. M. O'Hare. 2009. "Where Robots and Virtual Agents Meet." *International Journal of Social Robotics* 1 (1): 83–93.

Hong, B. B., E. Bohemia, R. Neubauer, and L. Santamaria. 2018. "Designing for Users: The Global Studio." In *DS 93: Proceedings of the 20th International Conference on Engineering and Product Design Education (E&PDE 2018), Dyson School of Engineering, Imperial College, London. 6th-7th September 2018*, 738–743.

Huang, W., I. Weber, and S. Vieweg. 2014. "Inferring Nationalities of Twitter Users and Studying Inter-National Linking." *ACM HyperText Conference*. https://works.bepress.com/vieweg/18/.

Idoughi, D., A. Seffah, and C. Kolski. 2012. "Adding User Experience into the Interactive Service Design Loop: A

Persona-Based Approach." *Behaviour & Information Technology* 31 (3): 287–303. doi:10.1080/0144929X.2011.563799.

Iizuka, S., E. Simo-Serra, and H. Ishikawa. 2017. "Globally and Locally Consistent Image Completion." *ACM Transactions on Graphics* 36 (4): 107:1–107:14. doi:10.1145/3072959.3073659.

Isola, P., J.-Y. Zhu, T. Zhou, and A. A. Efros. 2017. "Image-to-Image Translation with Conditional Adversarial Networks." In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Honolulu, HI. 5967–5976.

Jansen, B. J., J. O. Salminen, and S.-G. Jung. 2020. "Data-Driven Personas for Enhanced User Understanding: Combining Empathy with Rationality for Better Insights to Analytics." *Data and Information Management* 4 (1): 1–17. doi:10.2478/dim-2020-0005.

Jansen, A., M. Van Mechelen, and K. Slegers. 2017. "Personas and Behavioral Theories: A Case Study Using Self-Determination Theory to Construct Overweight Personas." In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2127–2136. doi:10.1145/3025453.3026003.

Jeffreys, H. 1998. *Theory of Probability*. 3rd ed. Oxford: Oxford University Press.

Jenkinson, A. 1994. "Beyond Segmentation." *Journal of Targeting, Measurement and Analysis for Marketing* 3 (1): 60–72.

Jung, S., J. Salminen, J. An, H. Kwak, and B. J. Jansen. 2018a. "Automatically Conceptualizing Social Media Analytics Data via Personas." In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2018), June 25*. International AAAI Conference on Web and Social Media (ICWSM 2018), San Francisco, California, USA.

Jung, S., J. Salminen, H. Kwak, J. An, and B. J. Jansen. 2018b. "Automatic Persona Generation (APG): A Rationale and Demonstration." In *Proceedings of the ACM, 2018 Conference on Human Information Interaction & Retrieval*, ACM, New Brunswick, NJ., 321–324.

Karras, T., S. Laine, and T. Aila. 2019. "A Style-Based Generator Architecture for Generative Adversarial Networks." In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4396–4405. doi:10.1109/CVPR.2019.00453.

King, A. J., A. J. Lazard, and S. R. White. 2020. "The Influence of Visual Complexity on Initial User Impressions: Testing the Persuasive Model of Web Design." *Behaviour & Information Technology* 39 (5): 497–510. doi:10.1080/0144929X.2019.1602167.

Lee, M. D. 2014. *Bayesian Cognitive Modeling: A Practical Course.* Cambridge: Cambridge University Press.

Lee, H.-Y., H.-Y. Tseng, J.-B. Huang, M. K. Singh, and M.-H. Yang. 2018. "Diverse Image-to-Image Translation via Disentangled Representations." In *Computer Vision – ECCV 2018. ECCV 2018 (Vol. 11205)*. Cham: Springer. doi:10.1007/978-3-030-01246-5_3.

Li, T., R. Qian, C. Dong, S. Liu, Q. Yan, W. Zhu, and L. Lin. 2018. "BeautyGAN: Instance-Level Facial Makeup Transfer with Deep Generative Adversarial Network." In *Proceedings of the 26th ACM International Conference on Multimedia*, 645–653. doi:10.1145/3240508.3240618.

Lin, C. H., C.-C. Chang, Y.-S. Chen, D.-C. Juan, W. Wei, and H.-T. Chen. 2019. "COCO-GAN: Generation by Parts via Conditional Coordinating." In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 4511–4520. doi:10.1109/ICCV.2019.00461.

Liu, Y., Z. Qin, T. Wan, and Z. Luo. 2018. "Auto-painter: Cartoon Image Generation from Sketch by Using Conditional Wasserstein Generative Adversarial Networks." *Neurocomputing* 311: 78–87. doi:10.1016/j.neucom.2018.05.045.

Liu, S., Y. Sun, D. Zhu, R. Bao, W. Wang, X. Shu, and S. Yan. 2017. "Face Aging with Contextual Generative Adversarial Nets." In *Proceedings of the 25th ACM International Conference on Multimedia*, 82–90. doi:10.1145/3123266.3123431.

Long, F. 2009. "Real or Imaginary: The Effectiveness of Using Personas in Product Design." In *Proceedings of the Irish Ergonomics Society Annual Conference*, Dublin, IR., 14.

Lu, Y., Y.-W. Tai, and C.-K. Tang. 2017. "Conditional Cyclegan for Attribute Guided Face Image Generation." *ArXiv Preprint ArXiv:1705.09966*.

Matthews, T., T. Judge, and S. Whittaker. 2012. "How Do Designers and User Experience Professionals Actually Perceive and Use Personas?" In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1219–1228. doi:10.1145/2207676.2208573.

McGinn, J. J., and N. Kotamraju. 2008. "Data-Driven Persona Development." In *ACM Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, New York, NY, USA, 1521–1524.

Neumann, A., C. Pyromallis, and B. Alexander. 2018. "Evolution of Images with Diversity and Constraints Using a Generative Adversarial Network." Cheng L., Leung A., Ozawa S. (eds). In *International Conference on Neural Information Processing*, 452–465.

Nie, D., R. Trullo, J. Lian, C. Petitjean, S. Ruan, Q. Wang, and D. Shen. 2017. "Medical Image Synthesis with Context-Aware Generative Adversarial Networks." Descoteaux M., Maier-Hein L., Franz A., Jannin P., Collins D., Duchesne S. (eds). In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 417–425.

Nielsen, L. 2019. *Personas—User Focused Design*. 2nd ed. Springer.

Nielsen, L., K. S. Hansen, J. Stage, and J. Billestrup. 2015. "A Template for Design Personas: Analysis of 47 Persona Descriptions from Danish Industries and Organizations." *International Journal of Sociotechnology and Knowledge Development* 7 (1): 45–61. doi:10.4018/ijskd.2015010104.

Nielsen, L., S.-G. Jung, J. An, J. Salminen, H. Kwak, and B. J. Jansen. 2017. "Who Are Your Users?: Comparing Media Professionals' Preconception of Users to Data-Driven Personas." In *Proceedings of the 29th Australian Conference on Computer-Human Interaction*, 602–606. doi:10.1145/3152771.3156178.

Nielsen, L., and K. Storgaard Hansen. 2014. "Personas Is Applicable: A Study on the Use of Personas in Denmark." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, New York, NY.. 1665–1674.

Özmen, M. U., and E. Yucel. 2019. "Handling of Online Information by Users: Evidence from TED Talks." *Behaviour & Information Technology* 38 (12): 1309–1323. doi:10.1080/0144929X.2019.1584244.

Palan, S., and C. Schitter. 2018. "Prolific. Ac—A Subject Pool for Online Experiments." *Journal of Behavioral and Experimental Finance* 17: 22–27.

Pitkänen, L., and J. Salminen. 2013. "Managing the Crowd: A Study on Videography Application." In *Proceedings of Applied Business and Entrepreneurship Association International (ABEAI), November*.

Probster, M., M. E. Haque, and N. Marsden. 2018. "Perceptions of Personas: The Role of Instructions." In *2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, 1–8. doi:10.1109/ICE.2018.8436339.

Pruitt, J., and T. Adlin. 2006. *The Persona Lifecycle: Keeping People in Mind Throughout Product Design*. 1st ed. Morgan Kaufmann.

Rönkkö, K. 2005. "An Empirical Study Demonstrating How Different Design Constraints, Project Organization and Contexts Limited the Utility of Personas." In *Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences – Volume 08*. doi:10.1109/HICSS.2005.85.

Rönkkö, K., M. Hellman, B. Kilander, and Y. Dittrich. 2004. "Personas Is Not Applicable: Local Remedies Interpreted in a Wider Context." In *Proceedings of the Eighth Conference on Participatory Design: Artful Integration: Interweaving Media, Materials and Practices – Volume 1*, 112–120. doi:10.1145/1011870.1011884.

Salimans, T., I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen. 2016. "Improved Techniques for Training GANs." In *Advances in Neural Information Processing Systems 29*, edited by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, 2234–2242. Curran Associates, Inc. http://papers.nips.cc/paper/6125-improved-techniques-for-training-gans.pdf.

Salminen, J., H. Almerekhi, P. Dey, and B. J. Jansen. 2018a. "Inter-Rater Agreement for Social Computing Studies." In *Proceedings of The Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS – 2018), October 15*. The Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS – 2018), Valencia, Spain.

Salminen, J., B. J. Jansen, J. An, H. Kwak, and S. Jung. 2018b. "Are Personas Done? Evaluating Their Usefulness in the Age of Digital Analytics." *Persona Studies* 4 (2): 47–65. doi:10.21153/psj2018vol4no2art737.

Salminen, J., B. J. Jansen, J. An, H. Kwak, and S.-G. Jung. 2019a. "Automatic Persona Generation for Online Content Creators: Conceptual Rationale and a Research Agenda." In *Personas—User Focused Design*, edited by L. Nielsen, 135–160. London: Springer. doi:10.1007/978-1-4471-7427-1_8.

Salminen, J., S. Jung, J. An, H. Kwak, L. Nielsen, and B. J. Jansen. 2019b. "Confusion and Information Triggered by Photos in Persona Profiles." *International Journal of Human-Computer Studies* 129: 1–14. doi:10.1016/j.ijhcs.2019.03.005.

Salminen, J., S.-G. Jung, and B. J. Jansen. 2019c. "Detecting Demographic Bias in Automatically Generated Personas." In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, LBW0122:1–LBW0122:6. doi:10.1145/3290607.3313034.

Salminen, J., S. G. Jung, and B. J. Jansen. 2019d. "The Future of Data-Driven Personas: A Marriage of Online Analytics

Numbers and Human Attributes." In *ICEIS 2019 – Proceedings of the 21st International Conference on Enterprise Information Systems*, 596–603. https://pennstate.pure.elsevier.com/en/publications/the-future-of-data-driven-personas-a-marriage-of-online-analytics.

Salminen, J., S.-G. Jung, J. M. Santos, and B. J. Jansen. 2019e. "The Effect of Smiling Pictures on Perceptions of Personas." In *UMAP'19 Adjunct: Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*. doi:10.1145/3314183.3324973.

Salminen, J., S.-G. Jung, J. M. Santos, and B. J. Jansen. 2019f. "Does a Smile Matter if the Person Is Not Real?: The Effect of a Smile and Stock Photos on Persona Perceptions." *International Journal of Human–Computer Interaction* 0 (0): 1–23. doi:10.1080/10447318.2019.1664068.

Salminen, J., H. Kwak, J. M. Santos, S.-G. Jung, J. An, and B. J. Jansen. 2018c. "Persona Perception Scale: Developing and Validating an Instrument for Human-Like Representations of Data." In *CHI'18 Extended Abstracts: CHI Conference on Human Factors in Computing Systems Extended Abstracts Proceedings*. CHI 2018 Extended Abstracts, Montréal, Canada. doi:10.1145/3170427.3188461.

Salminen, J., L. Nielsen, S.-G. Jung, J. An, H. Kwak, and B. J. Jansen. 2018d. "'Is More Better?': Impact of Multiple Photos on Perception of Persona Profiles." In *Proceedings of ACM CHI Conference on Human Factors in Computing Systems (CHI2018), April 21*. ACM CHI Conference on Human Factors in Computing Systems (CHI2018), ACM, Montréal, Canada. Paper 317.

Salminen, J., J. M. Santos, S.-G. Jung, M. Eslami, and B. J. Jansen. 2019g. "Persona Transparency: Analyzing the Impact of Explanations on Perceptions of Data-Driven Personas." *International Journal of Human–Computer Interaction* 0 (0): 1–13. doi:10.1080/10447318.2019.1688946.

Salminen, J., S. Şengün, H. Kwak, B. J. Jansen, J. An, S. Jung, S. Vieweg, and F. Harrell. 2017. "Generating Cultural Personas from Social Data: A Perspective of Middle Eastern Users." In *Proceedings of The Fourth International Symposium on Social Networks Analysis, Management and Security (SNAMS-2017)*. The Fourth International Symposium on Social Networks Analysis, Management and Security (SNAMS-2017), IEEE. Prague, Czech Republic, 1-8.

Salminen, J., S. Şengün, H. Kwak, B. J. Jansen, J. An, S. Jung, S. Vieweg, and F. Harrell. 2018e. "From 2,772 Segments to Five Personas: Summarizing a Diverse Online Audience by Generating Culturally Adapted Personas." *First Monday* 23 (6). doi:10.5210/fm.v23i6.8415.

Şengün, S. 2014. "A Semiotic Reading of Digital Avatars and Their Role of Uncertainty Reduction in Digital Communication." *Journal of Media Critiques* 1 (Special): 149–162.

Sengün, S. 2015. "Why Do I Fall for the Elf, When I am No Orc Myself? The Implıcatıons of Vırtual Avatars ın Dıgıtal Communıcatıon." *Comunicação e Sociedade* 27: 181–193.

Shmelkov, K., C. Schmid, and K. Alahari. 2018. "How Good Is My GAN?" In *Computer Vision – ECCV 2018*, edited by V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, 218–234. Springer International Publishing. doi:10.1007/978-3-030-01216-8_14.

Shmueli-Scheuer, M., T. Sandbank, D. Konopnicki, and O. P. Nakash. 2018. "Exploring the Universe of Egregious

Conversations in Chatbots." In *Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion*, ACM, New York, NY, Article 16.

Song, Y., D. Li, A. Wang, and H. Qi. 2019. "Talking Face Generation by Conditional Recurrent Adversarial Network." In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, Macao, China. 919–925.

Tan, W. R., C. S. Chan, H. E. Aguirre, and K. Tanaka. 2017. "ArtGAN: Artwork Synthesis with Conditional Categorical GANs." In *2017 IEEE International Conference on Image Processing (ICIP)*, IEEE. 3760–3764.

Weiss, J. K., and E. L. Cohen. 2019. "Clicking for Change: The Role of Empathy and Negative Affect on Engagement with a Charitable Social Media Campaign." *Behaviour & Information Technology* 38 (12): 1185–1193. doi:10.1080/0144929X.2019.1578827.

Wongpakaran, N., T. Wongpakaran, D. Wedding, and K. L. Gwet. 2013. "A Comparison of Cohen's Kappa and Gwet's AC1 when Calculating Inter-Rater Reliability Coefficients: A Study Conducted with Personality Disorder Samples." *BMC Medical Research Methodology* 13 (1): 61.

Yang, X., Y. Li, and S. Lyu. 2019. "Exposing Deep Fakes Using Inconsistent Head Poses." In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 8261–8265.

Yin, W., Y. Fu, L. Sigal, and X. Xue. 2017. "Semi-Latent GAN: Learning to Generate and Modify Facial Images from Attributes." *ArXiv:1704.02166 [Cs].* http://arxiv.org/abs/1704.02166.

Yuan, Y., H. Su, J. Liu, and G. Zeng. 2020. "Locally and Multiply Distorted Image Quality Assessment via Multi-Stage CNNs." *Information Processing & Management* 57 (4): 102175. doi:10.1016/j.ipm.2019.102175.

Zhang, X., H.-F. Brown, and A. Shankar. 2016. "Data-driven Personas: Constructing Archetypal Users with Clickstreams and User Telemetry." In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ACM, New York, NY, 5350–5359.

Zhang, R., P. Isola, A. A. Efros, E. Shechtman, and O. Wang. 2018. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric." In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 586–595. doi:10.1109/CVPR.2018.00068.

Zhao, R., Y. Xue, J. Cai, and Z. Gao. 2020. "Parsing Human Image by Fusing Semantic and Spatial Features: A Deep Learning Approach." *Information Processing & Management* 57 (6): 102306. doi:10.1016/j.ipm.2020.102306.

Zhou, M. X., W. Chen, Z. Xiao, H. Yang, T. Chi, and R. Williams. 2019a. "Getting Virtually Personal: Chatbots who Actively Listen to You and Infer Your Personality." In *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion*, ACM, New York, NY, 123–124.

Zhou, H., Y. Liu, Z. Liu, P. Luo, and X. Wang. 2019b. "Talking Face Generation by Adversarially Disentangled Audio-Visual Representation." *Proceedings of the AAAI Conference on Artificial Intelligence* 33 (01): 9299–9306. doi:10.1609/aaai.v33i01.33019299.