Transactions on Human-Computer Interaction

Volume 13 Issue 4

12-2021

Toxic Text in Personas: An Experiment on User Perceptions

Joni Salminen

Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha and Turku School of Economics at the University of Turku, Turku, Finland, jsalminen@hbku.edu.qa

Soon-gyo Jung

Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, sjung@hbku.edu.qa

João M. Santos

Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal, jmcsmsantos@gmail.com

Bernard J. Jansen

Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, jjansen@acm.org

Recommended Citation

Salminen, J., Jung, S.G., Santos, J. M., & Jansen, B. J. (2021). Toxic Text in Personas: An Experiment on User Perceptions. *AIS Transactions on Human-Computer Interaction*, *13*(4), pp. xx-xx.

DOI: 10.17705/1thci.00101

Available at http://aisel.aisnet.org/thci/vol13/iss4/X

Transactions on Human-Computer Interaction

Research Paper DOI: 10.17705/1thci.00101 ISSN: 1944-3900

Toxic Text in Personas: An Experiment on User Perceptions

Joni Salminen
Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar; and Turku School of Economics at the University of Turku, Turku, Finland

Soon-gyo Jung
Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha

João M. Santos
Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal

Bernard J. Jansen
Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha

Abstract:

When algorithms create personas from social media data, the personas can become noxious via automatically including toxic comments. To investigate how users perceive such personas, we conducted a 2 × 2 user experiment with 496 participants that showed participants toxic and non-toxic versions of data-driven personas. We found that participants gave higher credibility, likability, empathy, similarity, and willingness-to-use scores to non-toxic personas. Also, gender affected toxicity perceptions in that female toxic data-driven personas scored lower in likability, empathy, and similarity than their male counterparts. Female participants gave higher perceptions scores to non-toxic personas and lower scores to toxic personas than male participants. We discuss our results' implications for designing data-driven personas.

Keywords: Personas, Toxicity, Social Media, User Experiment.

Fiona Nah

1 Introduction

Personas, which originate from Cooper's Cooper (1999) work in the human-computer interaction (HCI) community, are fictitious people that describe core users or customers of a software system, product, or service. Organizations use personas to support user-centered activities (Anvari, Richards, Hitchens, & Tran, 2019) in software development, design, marketing, health informatics, and other fields (Aoyama, 2005; Holmgard, Green, Liapis, & Togelius, 2018; LeRouge, Ma, Sneha, & Tolle, 2013; Miaskiewicz & Kozar, 2011; Minichiello, Hood, & Harkness, 2018). Because personas aggregate users or customers under one "type", decision makers can consider the persona's needs and wants in design processes (Grudin, 2006; Pruitt & Grudin, 2003) and communicate these needs and wants to others in the organization (Blomquist & Arvola, 2002; Nielsen, 2002; Nielsen & Hansen, 2014).

Even though organizations widely use personas and become well established in research, manually creating personas suffers from several shortcomings (e.g., slowness, small sample sizes, poor objectivity, high costs, and unstable user behavior over time) (Chapman & Milham, 2006; Jung, Salminen, & Jansen, 2019; Salminen, Jansen, An, Kwak, & Jung, 2019a; Salminen, Jansen, An, Kwak, & Jung, 2018b). To address these challenges, researchers have developed methodologies for creating data-driven personas (An, Kwak, Jung, Salminen, & Jansen, 2018a; An, Kwak, Salminen, Jung, & Jansen 2018b; McGinn & Kotamraju, 2008; Molenaar, 2017; Salminen et al., 2018e; Zhang, Brown, & Shankar, 2016). The availability of social media user data and advances in data science algorithms and Web technologies have promulgated data-driven personas (DDPs) (Salminen, Guan, Chowdhury, & Jansen, 2020b). These DDPs typically contain various information about the users that they portray (see Figure 1), such as quotes from social media that users interpret as reflecting a persona's opinions and attitudes. Quotes represent essential information in persona profiles (Nielsen, Hansen, Stage, & Billestrup, 2015) and influence how persona users form impressions about a persona (Salminen et al., 2019b; Salminen, Jung, An, Kwak, & Jansen, 2018c; Salminen et al., 2018d).

Advances in persona creation and particularly the fact that many actors create DDPs from social media user data have increased the probability that toxic comments will appear in data-driven personas (Salminen et al., 2018c; Salminen et al., 2018d). "Toxic" refers to comments that one writes with harmful intent and that usually attack a person or group (Kocoń et al., 2021; Salminen et al., 2018a). Toxic comments pose a problem because algorithms automatically enrich DDPs with these social media comments (Mijač, Jadrić, & Ćukušić, 2018). Even when one relies on sophisticated algorithms (Pamungkas, Basile, & Patti, 2021), without human supervision, toxic quotes may appear in personas depending on the data source (Fortuna, Soler-Company, & Wanner, 2021). These toxic quotes may negatively affect . For instance, toxic quotes may:

- **Reduce user empathy towards a persona**—a key benefit associated with using personas for design in the first place (Nielsen, 2019).
- **Risk drawing users' attention away from more relevant information for the task at hand** (e.g., learning about the online audience segment (Salminen et al., 2019b)).
- **Contaminate a persona**. When algorithms create personas from user data that involves social media comments, the toxicity of these comments can pass into the persona profile.

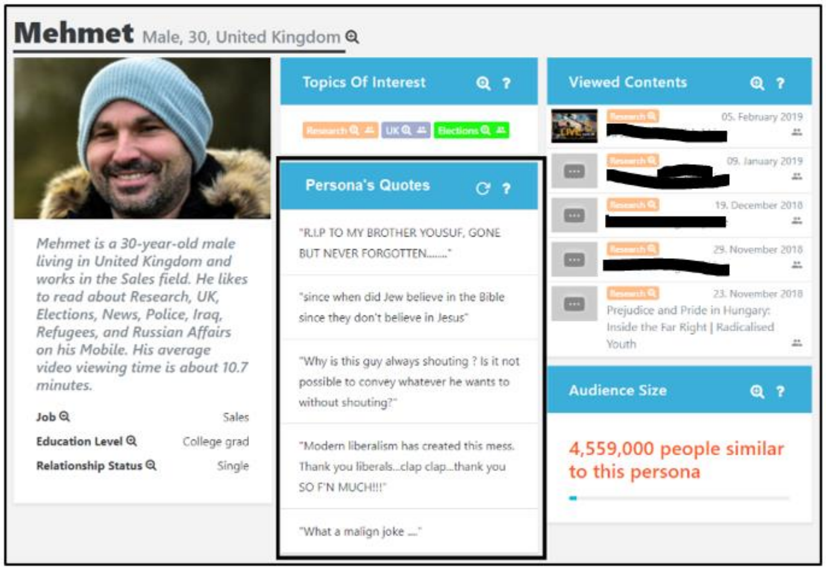


Figure 1. Persona Profiles Typically Contain Quotes that Describe the Persona's Attitudes and Opinions (See Black Box in the Middle)

The news vertical is one of the most prominent domains exposed to online toxicity. News and media channels face major toxic backlash due to reporting political, religious, and other controversial stories (Kittur, Chi, & Suh, 2009; Salminen et al., 2018a; Salminen, Sengün, Corporan, Jung, & Jansen, 2020). As such, news constitutes an extremely challenging domain for generating DDPs. Therefore, in this study, we particularly focus on the online news context, create DDPs from a news channel's data, and experiment how persona users perceive these personas when they include toxic and non-toxic quotes.

Creating DDPs from social media data differs drastically to creating them via traditional qualitative methods wherein creators have precise control over the content they select or create for personas. In DDPs, algorithms pick the quotes without considering the impact they may have on persona user perceptions. The choices that algorithms make pose a particularly vexing issue for personas generated for organizations whose content contains much toxicity. Examples, among others, include online news channels that toxic comments regularly target (Mejova, Zhang, Diakopoulos, & Castillo**,** 2014; Salminen et al., 2018a).

However, despite these possible negative side effects, research has not empirically investigated the impact that toxic social media quotes have on *persona user perceptions*. We require quantitative research to validate the impact that toxic persona quotes have on user perceptions since organizations increasingly automate persona creation (Mijač et al., 2018; Salminen et al., 2020b). In this study, we address this need by examining the following research question (RQ):

RQ: How do toxic text quotes affect how users perceive personas?

To address this question, we conducted an experimental user study in which we exposed participants to two DDPs: one with toxic quotes and the other with non-toxic quotes. We discuss our work's implications for persona creators and designers using personas generated from online social media data.

2 Related Literature

2.1 User Perceptions of (Data-driven) Personas

Researchers in HCI have repeatedly discovered that (Chapman & Milham, 2006; Friess, 2012; Matthews, Judge, & Whittaker, 2012). However, little quantitative research has examined how individuals perceive personas (Marsden & Pröbster, 2019; Pröbster, Hermann, & Marsden, 2019). Typically, previous studies have examined persona use via case studies (Faily & Flechais, 2011; Jansen, Van Mechelen, & Slegers, 2017), ethnography (Friess, 2012), usability (Long, 2009), or other qualitative means.

For example, Friess (2012) investigated persona adoption among designers. Long (2009) employed usability heuristics to measure persona use effectiveness. Anvari et al. (2019) and Anvari and Richards

(2016) studied how personas support learning outcomes. Researchers have conducted experimental evaluations more rarely; examples include LeRouge et al. (2013) and Ma and LeRouge (2007), who compared personas against user profiles, and Salminen et al. (2020), who compared personas against an analytics system.

Research has shown that user perceptions are critical for deploying and using personas in real organizations and use cases. Important perceptions that such work has mentioned include perceptions about credibility (Matthews et al., 2012), accuracy (Chapman & Milham, 2006), trust (Blomquist & Arvola, 2002), sense of immersion and (Nielsen, 2019), a persona's personality traits (Anvari et al., 2017; Anvari, Richards, Hitchens, & Babar, 2015), and a persona's perceived usefulness for the task at hand (Cooper, 1999; Salminen et al., 2018b). Also, prior research has established that persona perceptions vary individually and involve varied and subjective interpretations due to designers' different backgrounds and personal experiences (Hill et al., 2017; Marsden & Haag, 2016; Salminen et al., 2018d).

When researchers consider personas as targets of individuals' perceptions, they can logically contrast persona perception with *person* perception, a concept from social psychology (Grudin, 2006) that refers to a "general tendency to form impressions of other people" (Psychology Research and Reference, 2018)). The beliefs that individuals attribute to others can relate to looks, demographics, behaviors, dispositions, and other features (Ambady & Rosenthal, 1992). According to this premise, we define persona perception as a multi-dimensional perceptional construct that comprises fundamental beliefs that individuals intuitively associate with personas*.* Thus, the way in which people perceive and use personas depends on their idiosyncratic experiences.

Moreover, the above premise denotes a departure from the assumption that one should solely evaluate personas by measuring their accuracy in a technical sense (Chapman & Milham, 2006). Instead, we focus on how users perceive personas as people (Grudin, 2006). Nonetheless, we highlight the need for both validation types. Researchers should verify personas for accuracy (i.e., the extent to which they represent underlying data faithfully) (Chapman & Milham, 2006). We need studies that evaluate personas based on how individuals perceive different persona designs and types (Marsden & Haag, 2016). Prior work in the HCI community has examined both aspects to a limited degree, but we focus on the latter in this research.

Perceptions are *fundamental* when researchers use algorithms to create DDPs (Eslami, Krishna Kumaran, Sandvig, & Karahalios, 2018). Specifically, because researcher draw quotes from automatically from the content the persona has most engaged with (An et al., 2018a) (e.g., social media comments on news articles) the personas' toxicity is associated with the content's comments. The higher the share of toxic comments, the more likely the toxic quotes will "contaminate" the persona in the perceptual sense. Note that we presume here that "most users are *not* toxic", but some toxic comments (perhaps even one) can make the person appear toxic. This notion resembles the reputation-spoiling effect (Parveen, Jaafar, & Ainin, 2015), which refers to how one bad egg can spoil the whole basket, or, more generally, the halo effect, which describes how positive impressions about an individual in one area influences people to view the individual positively in other areas (e.g., as competent, skillfull, or successful) (Nisbett & Wilson, 1977). In contrast, people could view a toxic persona as all around "bad". While prior work has hinted at this risk associated with personas and their quotes (Salminen et al., 2018d), we could locate no previous empirical studies that has investigated such an effect.

As we discuss in Section 1, the toxic comments in DDPs can have several side effects, such as making it harder for users to relate to a persona to contaminating persona perceptions. Toxic comments may affect how designers perceive the persona, which may make the whole persona toxic in their eyes. The toxic comments can also draw attention away from other persona information designers should use for the decision-making task. Previous research has alluded to these effects (Nielsen et al., 2017; Salminen et al., 2019b; Salminen et al., 2018d), but we lack empirically rigorous evidence on either. Thus, researchers need to examine how one can detect toxic comments for personas created from social media data (Salminen et al., 2019a).

2.2 Hypothesis Development

We consider five relevant hypotheses for detecting changes in persona user perceptions when manipulating a persona's toxicity, which we explain below.

Likability refers to the tendency for someone to be liked by other people. It represents a notable concept in person perception literature (Reysen, 2005) since it affects the interactions between people. In particular, a persona's likeability may affect people's interest in the persona profile (Salminen, Jung, Santos, & Jansen, 2019d). Studies in social psychology have tended to show a negative association between likability and aggression and mental abuse (Kanekar, Bulsara, Duarte, & Kolsawalla, 1981). We expect that toxic quotes make a persona less likable as they demonstrate destructive attitudes. Hence, we propose:

H1: Users like toxic personas less than non-toxic personas.

Empathy refers to a feeling of understanding and compassion (Singer & Klimecki, 2014) and represents a crucial advantage when using personas for design (Marsden, Pröbster, Haque, & Hermann, 2017b; Nielsen & Hansen, 2014). Many believe that empathy drives motivation and purpose. In theory, empathetically understanding a persona helps designers to keep user needs in mind during design processes (Nielsen & Hansen, 2014). In social psychology, studies have shown that hostility (cf. toxicity) toward someone decreases one's empathy for that person (Wispé, 1986). We expect that there is a similar effect for personas, and toxic quotes reduce the sense of Empathy toward the persona. Hence, we propose:

H2: Users empathize less with toxic personas than with non-toxic personas.

Similarity refers to the experienced identification between a persona and its designer. Booth's (2008) findings imply that similarity to a persona may help individuals internalize its needs (Nielsen & Hansen, 2014; Pruitt & Grudin, 2003). Social psychologists have found that individuals are less likely to identify with people they perceive negatively (Rentsch & Woehr, 2004). Adopting this idea, we expect toxicity to reduce users' perceived similarity with personas. Hence, we propose:

H3: Users perceive toxic personas as less similar to themselves than non-toxic personas.

Willingness to use (WTU) refers to individuals' willingness to use a persona and represents a critical user perception since it dictates whether users actually use personas rather than forget about them after their creation (Rönkkö, Hellman, Kilander, & Dittrich, 2004). In the persona context, WTU relates to a person's willingness to learn more about a persona (Nielsen, Nielsen, Stage, & Billestrup, 2013), especially given the task at hand. Findings in social psychology show that hostile people drive people away rather than pull them in (Morris, Leung, & Iyengar, 2004). We expect that participants have less interest in learning about toxic personas. Hence, we propose:

H4: Users are less willing to use toxic personas than non-toxic personas.

Credibility refers to [what] and represents a notable challenge for persona use as users may perceive non-credible personas as abstract and unrealistic (Chapman & Milham, 2006; Matthews et al., 2012). Research has not established how toxicity affects personas' perceived credibility. However, toxicity may lessen the extent to which users perceive personas as credible due to the halo effect (NIsbett & Wilson, 1977). According to this idea, we hypothesize that users consider toxic personas less credible due to a general negative spillover effect. Hence, we propose:

H5: Users perceive toxic personas as less credible than non-toxic personas.

Overall, we can consider the five perceptions that the hypotheses address as *positive* impressions about personas (Salminen et al., 2020d). According to this idea, empirically derived persona perceptions can help designers design "good" (as in desirable and socially acceptable) personas. For example, a *high* willingness to use would be a desirable perception toward a persona as the persona would face less resistance from designers (Nielsen, 2004). As personas often face adoption barriers (Matthews et al., 2012; Rönkkö, 2005; Rönkkö et al., 2004), investigating users' attitudes towards different persona designs and information content can be impactful.

In addition to the above five hypotheses, we measure the effect of gender (both for personas and participants) as previous research on personas has shown that gender stereotypes may matter for persona perceptions (Hill et al., 2017; Marsden, Hermann, & Pröbster, 2017s). Notably, personas may reinforce existing gender stereotypes (Marsden & Haag, 2016) (Spiliotopoulos, Margaris, & Vassilakis, 2020). Users' perceptions might relate to a persona's gender, their own gender, or the interaction of the two (Salminen, Jung, Santos, Kamel, & Jansen, 2021). In this study, we expect the effect that personas' toxicity has on user perceptions to vary by persona gender and participant gender. Hence, we propose:

H6: The effect that persona toxicity has on user perceptions to varies by a) persona gender and b) user gender.

At first glance, the hypotheses may appear trivial in the sense that one might naturally expect people to negatively experience toxic personas. Nevertheless, having this preconception might result in fallacious implications for designing personas. Even though it would seem like toxicity would decrease positive user perceptions, for the sake of scientific inquiry, we should test the actual effects before drawing any conclusions because people often act and respond in unexpected ways (Harris, Spengler, & Gollery, 2016). Moreover, we have some specific grounds to expect alternative yet plausible outcomes. For example, users might still empathize with a toxic persona if they consider the persona to be misguided or in pain. In this case, users could disagree with a persona's views but still acknowledge that such views exist. Also, users could perceive toxic personas as more credible if they generally believe that social media involves many toxic people (and, thus, that toxic personas would be realistic). Moreover, if users agree with the content in a toxic message, they might feel more (and not less) similar to the toxic personas. Thus, in all, these hypotheses involve more complexity than first glance might suggest.

3 Methodology

First, we generated social media personas from the user data from an online news and media organization. Second, we manipulated these baseline personas' toxicity to create "toxic" and "non-toxic" personas. Third, we investigated how users reacted to these (non-)toxic personas. We explain our methodology in more detail in Sections 3.1 to 3.7 below.

3.1 Generation of Baseline Data-driven Personas

As we mention in Section X, data-driven personas (Jansen, Jung, Chowdhury, & Salminen, 2021a; Jansen, Jung, Chowdhury, & Salminen, 2021b; McGinn & Kotamraju, 2008) have become increasingly popular due to benefits relative to manual persona creation and due to the potential that "personified big data" provides (Stevenson & Mattson, 2019). While many methods to create DDPs exist, in this study, we employed automatic persona generation (APG), an interactive and state-of-the-art system to create DDPs (An et al., 2018a, 2018b) to generate personas from data that represented real online user audience segments. From the generated personas we generated, we chose two personas, one male and one female (see Figure 2). We chose these personas to allow for variation in gender but to keep age constant (Jiri is 30 and Kayla is 32).

A) Jiri (non-toxic version) B) Kayla (toxic version)

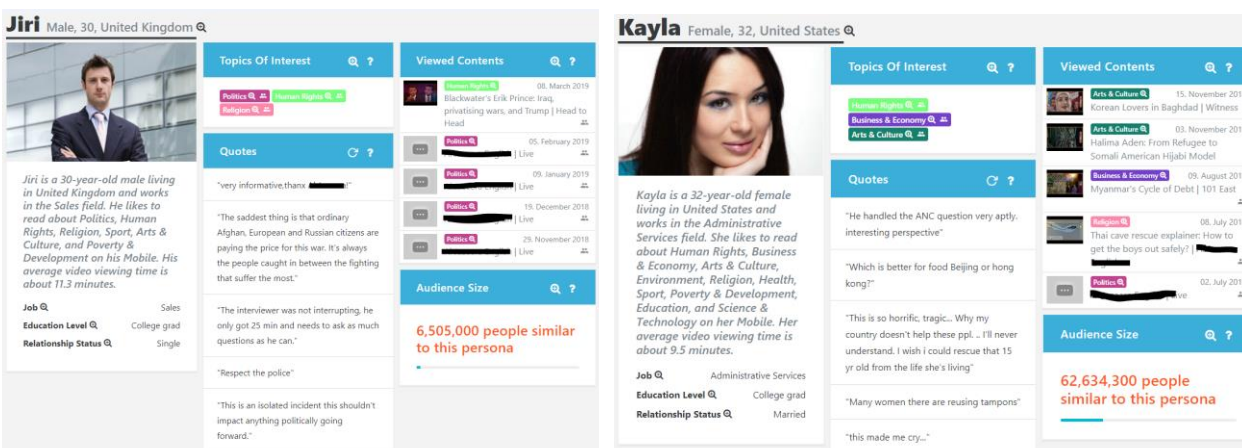


Figure 2. The Two Data-driven Personas We Chose for the Study

To provide background information to readers, we briefly explain how APG works to generate data-driven personas. Jung, Salminen, An, Kwak, and Jansen (2018a), Jung, Salminen, Kwak, An, and Jansen (2018b), and An et al. (2018a, 2018b) describe the algorithmic procedures in more technical detail. Overall, APG generates personas from quantitative user data via the following steps:

- **Step 1:** create an interaction matrix with videos as columns, demographic user groups as rows, and the view count of each group for each video as matrix elements.
- **Step 2:** apply non-negative matrix factorization (NMF) (Lee & Seung, 1999) to the interaction matrix to discern p latent video-viewing behaviors (where p is a hyper-parameter that we set).
- **Step 3:** choose the representative demographic attributes for each behavior by using weights from the NMF computation.

- **Step 4:** create the personas by enriching the representative demographic groups for each p personas with extra information, such as name, picture, topics of interest, and other attributes.

To create the personas we used in this research, we used data from an online news channel on YouTube that toxic commentators continuously attack (Salminen et al., 2018a). The attacks mainly target the channel itself, religious groups, and nationalities. Due to the high prevalence of toxic comments in this news channel's content, the DDPs often seem repulsive and toxic. We reached an agreement with the news and media organization to use the data for this research. We collected the dataset using the YouTube Analytics API[1] and according to YouTube Analytics' terms of service. The data contained no personally identifiable information about individual users apart from the username in the comments (which anyone can publicly see on the YouTube website). We do not show these usernames in the generated personas.

3.2 Toxic Quote Selection and Validation

We picked the (non-)toxic comments based on manually reviewing the comments in the dataset. One researcher identified candidate comments from both types (toxic/non-toxic) and asked the other members in the research team to independently assess a given comment as toxic or not. The comments combined xenophobia/racism, attacks against the media, and references to aggressive acts that the commenter would conduct (see Figure 3).

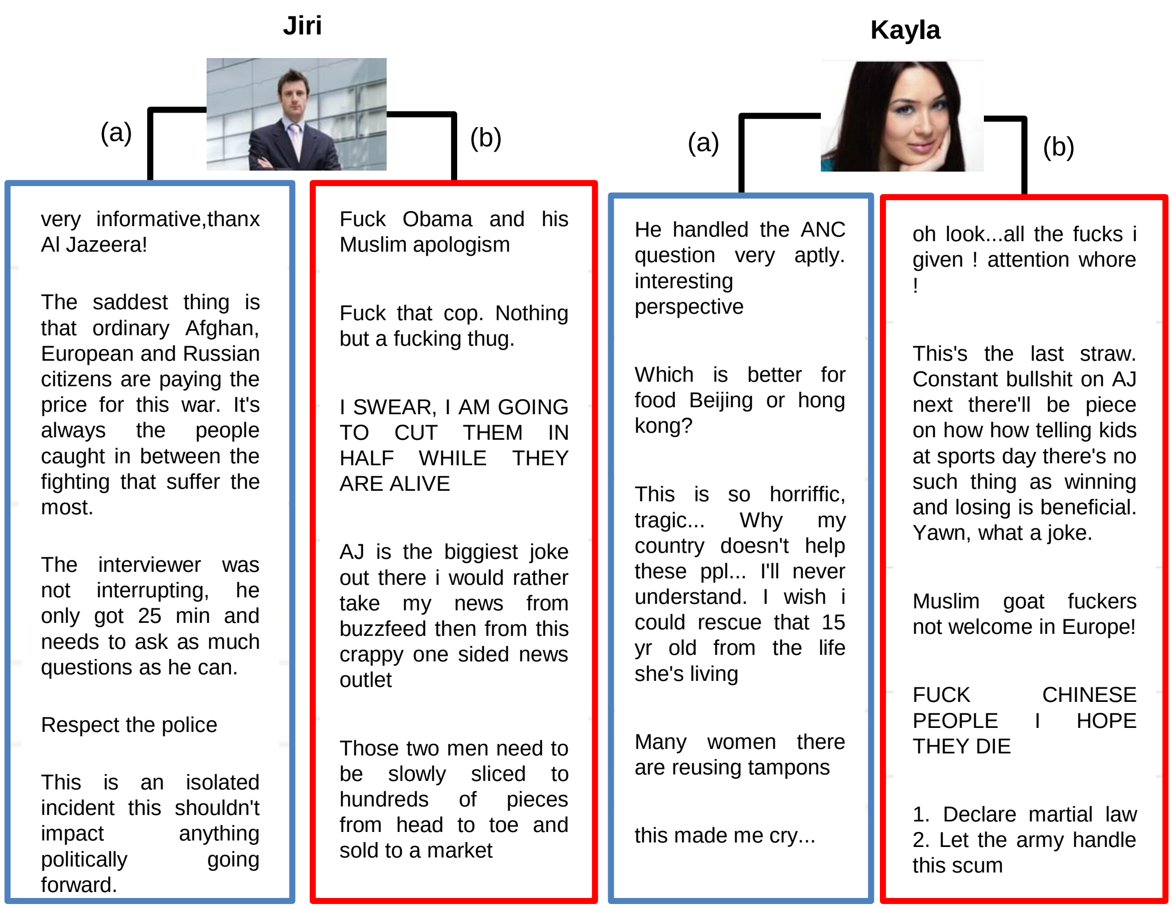


Figure 3. The Two Personas' Non-toxic Quotes (a) and Toxic Quotes (b)[2]

[1] See https://developers.google.com/youtube/analytics/

[2] We used real social media comments (although we masked the news company's name); thus, they reflect the harshness and toxicity that occurs in the wild. The supplementary material (see https://www.dropbox.com/s/xkwd2m66nx4joiy/supporting%20material%20-%20DIS21.zip?dl=0) shows the complete persona profiles. We first chose participants a toxic (or non-toxic) persona (Jiri/Kayla) and then a non-toxic (or toxic) persona (Kayla/Jiri).

We unanimously agreed on the (non-)toxicity of the candidate comments. Since one can consider toxicity a subjective variable, we conducted a manipulation check to ascertain the validity of the chosen comments (Hoewe, 2017). We conducted the manipulation check via a crowdsourcing platform (Appen[3]). Specifically, we showed crowd workers each comment and asked them if they found it toxic or not. For the task, we defined toxic as follows: "A toxic comment contains hostile or abusive content". Table 1 shows that that the crowd workers labeled presumably toxic comments as toxic (42.3%) and non-toxic comments as non-toxic (48.7%). We can attribute the slight deviations from expected values to individual differences in online hate interpretation (Salminen, Veronesi, Almerekhi, Jung, & Jansen, 2018f). Nonetheless, the results overwhelmingly support the comments we chose. Therefore, we proceeded to manually insert the comments into the persona profiles.

Table 1. Manipulation Check Results (N = 600 Crowd Ratings)

		Observed	
		Toxic	Non-toxic
Expected	Toxic	**254 (42.3%)**	46 (7.7%)
	Non-toxic	8 (1.3%)	**292 (48.7%)**

Note: percentage (%) indicates the share of ratings from the total ratings
The ratios of expected and observed toxic/non-toxic comments indicate that the toxicity levels in crowd ratings matched the expectations. Bolded cells indicate values that we expected to be close to 50%.

3.3 Persona Treatments

We created two toxic personas that contained toxic quotes and two non-toxic personas that did not contain toxic quotes. Overall, we created two versions of two personas, which resulted in four persona treatments in total: 1) non-toxic persona 1 ("Jiri"), 2) toxic persona 1, 3) non-toxic persona 2 ("Kayla"), and 4) toxic persona 2 (see Figure 2). We kept all information except comment toxicity, name, and gender identical in the persona profiles. By varying the gender (male/female), we could investigate gender-specific differences that, in general, the person perception literature has found to have an effect on user perceptions (Hill et al., 2017; Marsden & Haag, 2016). Also, this variation comes at almost no "cost" because a 2 × 2 design always requires two different personas.

3.4 Experiment Design

We placed the created personas in a 2 × 2 within-subject experiment design. We placed the four treatments into four sequences that we created using the survey software. We randomly distributed participants between these four sequences in a way that ensured each sequence contain nearly an equal number of participants. These choices followed the standard practices in experiment design (Brooks, 2012) that researchers have proposed for mitigating possible order and learning effects:

- **Sequence 1:** toxic Jiri → non-toxic Kayla (122 participants)
- **Sequence 2:** non-toxic Jiri → toxic Kayla (128 participants)
- **Sequence 3:** toxic Kayla → non-toxic Jiri (122 participants), and
- **Sequence 4:** non-toxic Kayla → toxic Jiri (124 participants).

We provide detailed information on the participants in Section 3.6.

3.5 Measurement Items

To measure the perceptions in our hypotheses, we used the persona perception scale (PPS), an instrument designed to gauge (Salminen et al., 2020d). Researchers have previously deployed the PPS in similar persona experiments, such as to test the effect that smiling images (Salminen et al., 2019d) and explanations in DDPs (Salminen, Santos, Jung, Eslami, & Jansen, 2019e) have on user perceptions. Salminen et al. (2020d) validated the PPS, although we also conducted a separate validation analysis here (see Section 4.4). We measured four persona perceptions from the PPS with 16 items in total (see Table 2).

[3] https://www.appen.com

Table 2. Survey Constructs and Items (Salminen et al., 2020d)

Construct	Reflecting…	Items
Likability	…how likable the persona feels.	I find this persona likable. I could be friends with this persona. This persona feels like someone I could spend time with. This persona is interesting.
Empathy	…how well the individual emotionally relates to the persona.	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.
Similarity	…how similar the individual feels the persona is to him/her.	This persona feels similar to me. The persona and I think alike. The persona and I share similar interests. I believe I would agree with this persona on most matters.
WTU	…how much the individual is interested in learning more about the persona.	I would make use of this persona in my task [of creating a YouTube video]. I would like to know more about this persona. I can imagine ways to make use of the persona information in my task [of creating the YouTube video]. This persona would improve my ability to make decisions about the customers it describes.
Credibility	…how realistic the persona appears.	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.
We measured the items on a seven-point Likert scale from "strongly disagree" to "strongly agree". The total score of a construct is the average of the scores of its items.		

3.6 Participant Recruitment and Information

After creating the four treatments, we created four surveys that corresponded to the counterbalanced experiment flows using an online survey-creation tool. To collect data, we recruited participants using Prolific[4], an online survey platform. Previous research, such as persona user studies (Salminen et al., 2019d, 2019e), has successfully applied Prolific to gauge how people perceive various topics (Palan & Schitter, 2018). Furthermore, research has found data collected from Prolific to have suitable reliability for research purposes (Palan & Schitter, 2018; Peer, Brandimarte, Samat, & Acquisti, 2017). We applied the following sampling criteria in Prolific: age: 23 to 50; nationality: United Kingdom (UK); minimum education level: undergraduate; and exclusion: did not participate in any other sequence.

With these criteria, we identified 5,173 eligible participants. We collected 130 answers for each flow (i.e., 4 × 125 = 520 answers in total) as we considered this number high enough for statistical analysis while remaining in our data-collection budget. We collected the data sequentially and used the platform's custom blacklist feature to exclude participation in more than one flow per user ID. Participants were 36.5 years (SD = 7.3) old on average, and 66.5 percent were female. All came from the UK. We focused on U.K. nationals since we wanted to 1) ensure that the participants could speak English fluently and 2) mitigate the impact of cultural variability that can affect how people interpret toxicity (Salminen et al., 2018f) and perceive personas (Putnam, Kolko, & Wood, 2012).

3.7 Survey Flow

First, each participant saw an introduction to the survey that explained what the survey focused on and personas ("a persona is defined as a fictive person describing a specific customer group"). Second, we explained the content in the shown persona profiles and how we retrieved it. Third, we explained the task ("Imagine that you are creating a YouTube video for the target group that the persona you will be shown next describes"). Fourth, we showed each participant one of the four treatments and asked them to review the information carefully and fill in the PPS questionnaire. To maintain ethical standards (see Appendix 1),

[4] https://prolific.ac

we informed the participants beforehand that the personas they saw contained some explicit content. Participants participated in the experiment on a voluntarily basis and could stop at any time.

4 Results

4.1 Data Processing and Analysis Procedure

After obtaining the answers to the survey, we checked their quality based on an included 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 to answer the attention check correctly. Thus, we removed these answers from the data. Also, the Prolific system timed out five participants. Therefore, we ended up with 496 qualified participants (95.4% of the total answers). For the statistical analysis, we applied repeated-measures mixed MANOVA (Hair, Black, Babin, & Anderson, 2009) to determine whether the toxic measurements significantly differed between the toxic and non-toxic conditions and whether the persona’s gender influenced these differences. We included the participants’ gender as a between-subjects dummy variable.

4.2 Validity Analysis

Although earlier research has validated the PPS instrument (Salminen et al., 2020d), we nonetheless performed an independent validation using the study sample. The results (see Table 3) showed satisfactory convergent validity such that the average variance extracted (AVE) of all constructs exceed 0.50 (Hair et al., 2009). Moreover, we found satisfactory discriminant validity, which requires that all pairs of factors’ square root of the AVE be equal to or greater than the correlations between these factors and cumulatively that the AVE for a given factor be greater than its maximum shared variance (MSV) and average shared variance (ASV) (Fornell & Larcker, 1981). Finally, we found satisfactory reliability as the composite reliability (CR) indicator exceeded the 0.7 threshold (Fornell & Larcker, 1981).

Table 3. Model Revalidation

					Correlations			
	CR	AVE	MSV	ASV	WTU	Credibility	Clarity	Likability
WTU	0.926	0.757	0.560	0.397	0.870			
Credibility	0.836	0.565	0.445	0.359	0.631	0.751		
Clarity	0.751	0.513	0.234	0.203	0.484	0.484	0.716	
Likability	0.979	0.938	0.560	0.381	0.748	0.667	0.374	0.969
Note: the diagonal of the correlation’s matrix indicates the square root of the AVE.								

4.3 Correlations

We computed the correlations for all variables under study (see Table 4). To do so, we used the mean for both toxic and non-toxic personas. As expected, all dependent variables exhibited significant correlations between them.

Table 4. Pearson Correlations for the Dependent Variables

Variable	Likability	Empathy	Similarity	WTU	Credibility
Likability	1	0.685***	0.760***	0.581***	0.496***
Empathy		1	0.662***	0.562***	0.470***
Similarity			1	0.509***	0.358***
WTU				1	0.438***
Credibility					1
Note: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$					

4.4 Hypotheses Testing

We began by analyzing the multivariate tests to identify whether any independent variable exhibited significant effects concerning any dependent variable. Researchers typically use the multivariate test before proceeding into the univariate analysis as it deflates type I errors by considering the existing correlations between the dependent variables (Hair et al., 2009). All effects were significant except the three-way interaction ($F(5, 481) = 0.635$, $p = 0.673$), which indicates that we could analyze them with follow-up univariate ANOVAs. We show the results in Table 5.

Table 5. Multivariate Tests (df(error) = 1(481))

Independent variable	Pillai's Trace	F	η2p	p-value
Toxicity	**0.743**	**277.944**	**0.743**	**< 0.001**
Toxicity x persona gender	**0.064**	**6.587**	**0.064**	**< 0.001**
Toxicity x participant gender	**0.069**	**7.174**	**0.069**	**< 0.001**
Toxicity x persona gender x participant gender	0.007	0.635	0.007	0.673
Note: significant results bolded (a = 0.05).				

We next proceeded into the univariate component of the analysis. The results in Table 6 show significant ($p < 0.001$) differences regarding toxic and non-toxic personas for all perceptions. The non-toxic personas scored significantly higher on all the tested perceptions (see Figure 4). Therefore, we found support for all five hypotheses.

Table 6. Univariate Tests for Within-subjects Effects and Interaction Terms with Between-subjects Effects (df(error) = 1(485))

Independent variable	Dependent variable	F	η2p	p-value
Toxicity	**Likability**	**1376.717**	**0.739**	**<0.001**
	Empathy	**852.115**	**0.637**	**<0.001**
	Similarity	**789.140**	**0.619**	**<0.001**
	Willingness to use	**504.575**	**0.510**	**<0.001**
	Credibility	**206.484**	**0.299**	**<0.001**
Toxicity × persona gender	**Likability**	**23.509**	**0.046**	**<0.001**
	Empathy	**6.558**	**0.013**	**0.011**
	Similarity	**12.096**	**0.024**	**0.001**
	Willingness to use	3.308	0.007	0.070
	Credibility	0.080	0.000	0.777
Toxicity × participant gender	**Likability**	**26.560**	**0.052**	**<0.001**
	Empathy	**15.045**	**0.030**	**<0.001**
	Similarity	**19.101**	**0.038**	**<0.001**
	Willingness to use	**29.118**	**0.057**	**<0.001**
	Credibility	**4.498**	**0.009**	**0.034**
Toxicity x persona gender x participant gender	Likability	1.174	0.002	0.279
	Empathy	0.027	0.000	0.869
	Similarity	0.977	0.002	0.324
	Willingness to use	0.216	0.000	0.642
	Credibility	0.888	0.002	0.346
Note: significant results bolded (a = 0.05)				

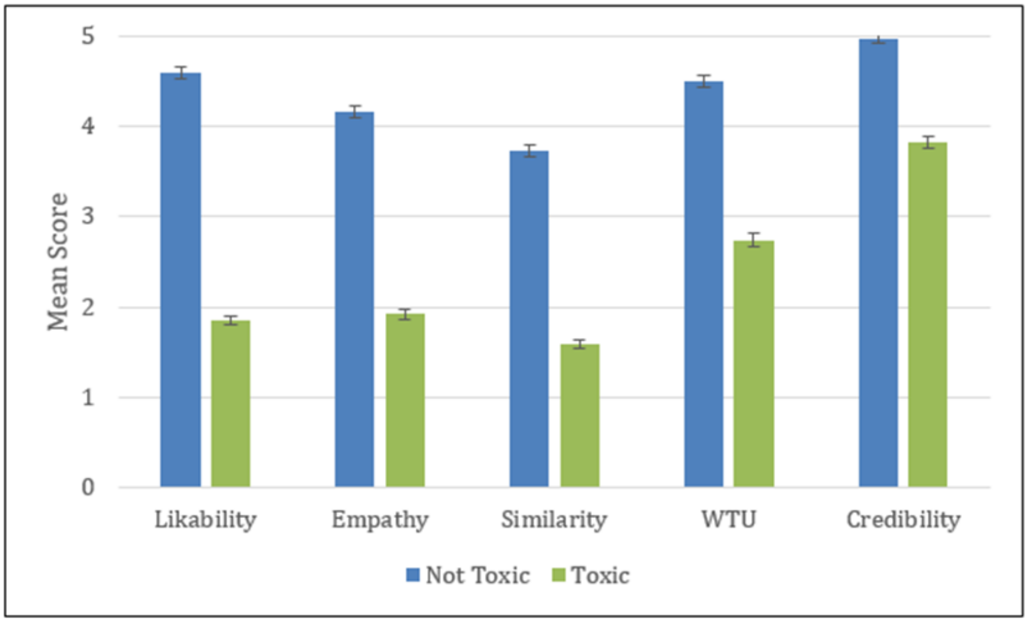


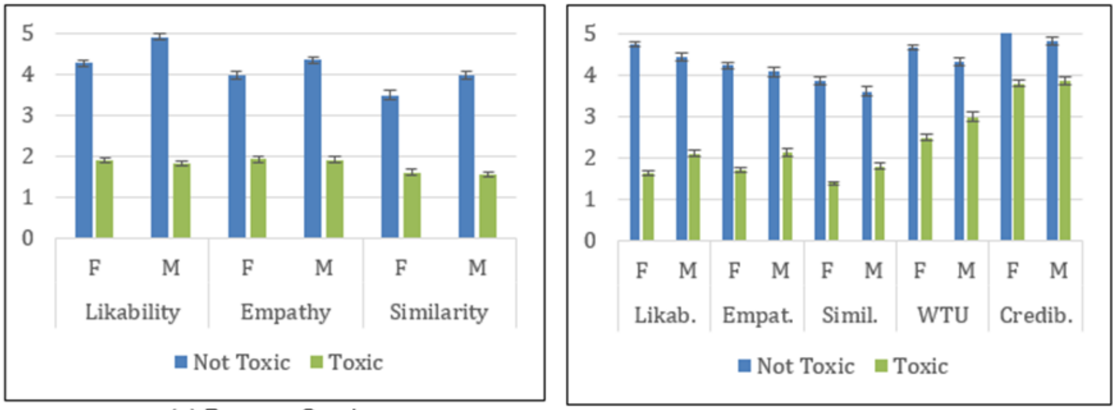
Figure 4. Differences in Perception Scores between Toxic and Non-toxic Personas[5]

Participants liked toxic personas (M = 1.86) less than non-toxic personas (M = 4.59), $F(1, 485) = 1376.71$, $p < 0.001$. Therefore, we found support for H1. Participants empathized with toxic personas (M = 1.92) less than with non-toxic personas (M = 4.16), $F(1, 485) = 852.11$, $p < 0.001$. Therefore, we found support for H2. Users perceived toxic personas as less similar to themselves (M = 1.59) than non-toxic personas (M = 3.73), $F(1, 485) = 789.14$, $p < 0.001$. Therefore, we found support for H3. Users were less willing to use toxic personas (M = 2.74) than non-toxic personas (M = 4.50), $F(1, 485) = 504.56$, $p < 0.001$. Therefore, we found support for H4. Users found toxic personas (M = 3.83) less credible than non-toxic personas (M = 4.97), $F(1, 485) = 206.48$, $p < 0.001$. Therefore, we found support for H5.

The scores for persona perceptions, when all other variables remained constant, increased as follows when comparing toxic to non-toxic personas: likability (+146.78%, $p < 0.001$), similarity (+134.59%, $p < 0.001$), empathy (+116.67%, $p < 0.001$), WTU (+64.23%, $p < 0.001$), and credibility (+29.77%, $p < 0.001$). Thus, the scores for likability, similarity, and empathy were more than double for non-toxic personas relative to toxic personas. In turn, WTU and credibility exhibited a more modest but still significant increase.

Regarding H6, we found two interaction effects regarding the gender variables—one for persona gender and the other for participant gender (see Figure 5). First, regarding the toxicity x persona gender interaction: female toxic personas scored significantly less on likability ($F(1, 485) = 23.509$, $p < 0.001$), empathy ($F(1, 485) = 6.558$, $p < 0.05$), and similarity ($F(1, 485) = 12.096$, $p < 0.01$) than the male toxic persona counterpart (see Figure 5a). These results indicate that toxicity negatively impacts both personas but that it more negatively impacts the female persona. Consistent with this finding, the likability score of the male persona had the highest increase (+169.78%): it went from M = 1.82 (toxic) to M = 4.91 (non-toxic). The similarity of the male persona had the second highest increase (+154.49%): the average score went from M = 1.56 (toxic) to M = 3.97 (non-toxic).

[5] All differences were significant ($p < .001$). Error bars indicate standard error.



(a) Persona Gender

(b) Participant Gender

Figure 5. Estimated PPS Scores by (a) Persona Gender and (b) Participant Gender (F = Female, M = Male)[6]

A second gender effect was the interaction between toxicity × participant gender, which shows that female participants tended to give higher scores to non-toxic personas and lower scores to toxic personas than male participants. In other words, females tended to exhibit more polarized reactions when contrasting toxic and non-toxic personas (see Figure 5b). Most notably, WTU decreased by 46.68 percent (i.e., from M = 4.67 (non-toxic) to M = 2.49 (toxic)) for female participants, whereas it decreased by only 32.64 percent (i.e., from M = 4.32 (non-toxic) to M = 2.91 (toxic)) for male participants.

Other persona perceptions decreased even more among female participants: likability decreased by 64.6 percent (12.3% more than for males), similarity decreased by 64.22 percent (13.99% more than for males), Empathy decreases by 59.82% (12.21% more than for males), and WTU decreased by 46.62 percent (15.79% more than for males). The difference between these perceptions was significantly smaller for the male participants than for the female participants, which indicates that persona toxicity affected female users more than male users.

To investigate this effect further, we conducted a cross-interaction analysis for persona gender x participant gender to analyze how gender correspondence affected user perceptions. We found that this interaction lacked significance (Pillai's T = 0.007, F(5, 481) = 0.673), which indicates the absence of a "gender-matching" effect. .

5 Discussion

5.1 Research Contribution

Our empirical analysis confirms the qualitative suggestions in the HCI literature that toxic comments affect how users perceive personas (Salminen et al., 2019a; Salminen, Jung, & Jansen, 2019c). Among our most significant findings, we found that WTU decreased for toxic personas. Designers need to consider this finding since, if they do not want to use the personas they create, persona creation becomes a futile exercise (Rönkkö et al., 2004). In particular, we found that users perceive personas without toxic quotes more favorably, which suggests that filtering the toxic comments could be helpful when creating personas from social media data with many toxic comments. This result held for both males and females, although female participants experienced toxic persona quotes more strongly.

This study offers three main takeaways:

1) An increase in toxicity in a persona's comments results in a decrease in all the measured persona perceptions
2) The increase in toxicity affected likability, similarity, and empathy the most.
3) The increase willingness to use and credibility to a lesser degree.

[6] We show only statistically significant differences. Error bars indicate standard error.

We find it logical that the affective dimensions of persona perception (likability, similarity, and empathy) decreased the most with the toxic quotes since these perceptions deal with how warmly a person thinks about a persona. Participants did not think of toxic personas very warmly.

In other words, individuals might still perceive toxic personas as (somewhat) interesting and credible. This finding is associated with the idea of holistic personas (Anvari & Tran, 2013), which postulates that persona perception holds subtle patterns (e.g., one can still consider disliked personas interesting for design). The weaker effect may indicate that some participants believed toxicity formed part of the persona's character. Overall, 23.6 percent of the participants found the toxic personas to be more or equally credible than non-toxic personas.

Finally, gender played a role in toxicity perceptions in several ways. The toxic female personas scored significantly less on likability, empathy, and similarity than the toxic male personas, which suggests that toxicity harms perceptions about female personas more than perceptions about male personas. Also, female participants tended to give higher scores to non-toxic personas and lower scores to toxic personas when compared to male participants. That is, toxicity affected female participants' perceptions more strongly than male participants' perceptions. These results suggest gendered thinking among the participants, which concurs with previous research (Hill et al., 2017). In practice, one cannot easily mitigate gendered effects in personas' design since they are ingrained into individuals' thinking about gender in general (Marsden & Haag, 2016).

5.2 Design Implications

As for practical implications for persona design and system development, the negative effect that toxic quotes had on persona perceptions implies that DDP developers should give special attention to the toxicity of persona quotes when using algorithms to generate DDPs from social media data. In particular, they need to exercise caution when [doing what?].

While we recommend applying toxicity *detection* when creating personas whose quotes originate from social media, we do not mean to imply that designers should *delete* toxic quotes from persona profiles by default. One cannot deal with toxicity simply by applying blind censorship. First, one needs to address whether filtering toxicity removes individuals' freedom of expression (Davidson, Warmsley, Macy, & Weber, 2017). That is, do users prone to toxicity have the *right* to be represented in online news personas? Second, one needs to consider accuracy; if people make many toxic comments, should not the journalists using personas to understand their audiences be aware of such comments? In other words, do journalists not have the right to know the dark side of their audience?

In a broader sense, the above discussion links to a more prominent theme in efforts to design computational systems. That is, how do we use the toxicity classification scores in real systems? Previous HCI research contains surprisingly little discussion on this matter even though researchers recognize toxicity as a serious issue in many computing systems (Fortuna & Nunes, 2018; Sengün, Salminen, Mawhorter, Jung, & Jansen, 2019; Türkay, Formosa, Adinolf, Cuthbert, & Altizer, 2020). Therefore, we need community efforts and discussion on the normative basis of toxicity management in computing systems: where does the HCI community stand on these matters?

Research also needs to pay more attention to DDP design principles. Namely, a "purist" approach to data-driven personas would claim that we should present all the truth and nothing but the truth in the personas we create. However, does toxicity filtering present a deviation from this rule? Does this rule matter in the first place? While we do not claim definite answers to these questions, we raise them as discussion points here because the HCI research community must define a position. Under which circumstances, if any, is it possible to deviate from the principle that data should drive personas? Toxic content and marginalized (fringe) personas present at least two ample cases where this question becomes practically relevant (Goodman-Deane et al., 2018; Salminen, Froneman, Jung, Chowdhury, & Jansen, 2020a). Namely, if data is centralized around, say, young White people, should we present no Black/Asian/elderly personas? Thus, we can see that the design tradeoffs require further thought; each direction offers pros and cons.

Some individuals who advocate for quantitative persona creation might argue that truthfulness to data—objectivity—outweighs improvements to user perceptions in importance (Chapman & Milham, 2006). Therefore, we find ourselves navigating these two conflicting needs: 1) the need to create truthful personas and 2) the need to create non-harmful personas. From a design point of view, designers face a dilemma in that, if they make personas "user friendly" by only incorporating non-toxic comments, they steer away from social media as it appears in reality. Withholding information on user attitudes can result in information bubbles and echo chambers (Del Vicario et al., 2016). On the one hand, the design goal should not be "make likable personas". On the other hand, (inadvertently) designing harmful personas can increase resistance to personas' use for design.

Based on our findings, giving users options to filter toxic quotes does seem like an essential feature for interactive persona systems. While designing such features goes beyond our scope here, we demonstrate a possible implementation in Figure 6.

Figure 6. Hate Filter Functionality[7]

Beyond the immediate context of personas in HCI, our research implications deal with how people perceive online user profiles with the implication that toxic comments in a user profile lead to an adverse effects on user perceptions. Here, we investigated a set of perceptions inspired by the nascent persona perception literature, itself inspired by the notion of person perception in social psychology. Further studies could measure other perceptions about toxicity impact how users perceive personas (and online profiles in general).

5.3 Limitations and Future Research Directions

The study has some limitations. First, we focused on persona quotes when manipulating toxicity. However, prior studies have shown that persona pictures also influence user perceptions (Hill et al., 2017; Long, 2009; Nieters, Ivaturi, & Ahmed, 2007; Salminen et al., 2018d). Therefore, future research could investigate the joint and separate impact that pictures have on user perceptions with quotes to better dissect the role that various persona information plays in how end users interpret personas overall. Here, we considered the persona to be "toxic" if (s)he had toxic quotes, but researchers could experiment with visual imagery (e.g., angry faces) to create similar conditions in contexts that reach beyond social media (e.g., researchers could test "unhappy customers" or "angry software users"). Further work in this area could explain in more detail what makes a persona toxic in end users' eyes and, thereby, increase our theoretical knowledge about persona-user interaction.

Second, we investigated only completely toxic versus completely non-toxic personas. Future studies could persona designs that contain both toxic and non-toxic quotes to varying degrees. Such research could determine a "toxicity breaking point" at which toxicity "takes over" users' perceptions and tilts them in a negative direction. For example, a smaller ratio of toxic quotes could possibly not trigger a significant backlash in user perceptions if users perceived a persona as a "real person" with negative and positive dimensionality. In polarized contexts such as politics, it would be interesting to test how personas of the opposite political spectrum would possibly help defuse conflicts, increase understanding, and build bridges between people.

Third, experiments dealing with sensitive topics such as online toxicity may be associated with social desirability bias (Fisher, 1993). While such studies may always potentially contain social desirability bias, we would argue that the risk becomes more negligible in a platform-mediated anonymous remote user study setting (such as our study) relative to an interview study where subjects directly interact with researchers because, in the former, subjects anonymously provide answers and do not socially interact with researchers in a way that might require them to obfuscate their true opinions. Although we could not find any specific study focused on social desirability bias in platform-mediated remote user studies (in fact, we find this research question itself interesting), we nonetheless believe that the lack of social interaction between

[7] By enabling or disabling the toggle, users can choose to see or not see toxic quotes in the generated persona profile.

subjects and researchers creates a setting where social desirability bias poses less concern (i.e., people feel confident in expressing their true opinions because they do not fear that others will morally judge them).

Fourth, while interestingness (item 4) could be associated with likability, cases may exist where one could characterize even a repulsive entity as interesting. Nonetheless, the AVE, AVS, CR, and MSV values in Table 4 indicate that likability behaved like a valid independent construct. Fifth, we focused on U.K. nationals in this study. The internationally high prevalence of toxicity means we need to understand social media users' toxicity beyond Western contexts (Chowdhury et al., 2020). Therefore, we encourage researchers to conduct cross-cultural studies that examine users' toxicity perceptions. Sixth, we measured toxicity as a dichotomous variable for parsimony (i.e., maintaining a 2 × 2 experimental design). However, researchers have observed toxicity perceptions to involve more fine-grained levels (Chatzakou et al., 2017). Thus, future studies could consider toxicity perception as a non-binary variable and perhaps measure it via a Likert scale. Finally, we only measured perceptions but not behavioral interactions with the (non-)toxic personas, which represents an exciting research direction.

6 Conclusion

We obtained empirical results show that toxic quotes in DDPs affect various persona user perceptions negatively. Toxic DDPs appeared less likable, less similar to users, and less empathetic. Participants also noted less willingness to learn more about toxic DDPs. Non-toxic DDPs, in turn, received a higher level of positive impressions from users concerning credibility, likability, empathy, similarity, and willingness to use. Our results imply that removing toxic quotes from persona profiles positively affects user perceptions. To avoid interference with the DDPs' truthfulness, we suggest that designers give persona users a choice in data-driven persona UI about whether to apply toxicity filtering or not.

References

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*(2), 256-274.

An, J., Kwak, H., Jung, S., Salminen, J., & Jansen, B. J. (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).

An, J., Kwak, H., Salminen, J., Jung, S., & Jansen, B. J. (2018b). Imaginary people representing real numbers: Generating personas from online social media data. *ACM Transactions on the Web, 12*(4), 1-26.

Anvari, F., & Richards, D. (2016). A method to identify talented aspiring designers in use of personas with personality. In L. A. Maciaszek & J. Filipe (Eds.), *Evaluation of novel approaches to software engineering* (vol. 599, pp. 40-61). Berlin: Springer.

Anvari, F., Richards, D., Hitchens, M., & Babar, M. A. (2015). Effectiveness of persona with personality traits on conceptual design. In *Proceedings of the 37th International Conference on Software Engineering.*

Anvari, F., Richards, D., Hitchens, M., Babar, M. A., Tran, H. M. T., & Busch, P. (2017). An empirical investigation of the influence of persona with personality traits on conceptual design. *Journal of Systems and Software*, *134*, 324-339.

Anvari, F., Richards, D., Hitchens, M., & Tran, H. M. T. (2019). Teaching user centered conceptual design using cross-cultural personas and peer reviews for a large cohort of students. In *Proceedings of the IEEE/ACM 41st International Conference on Software Engineering.*

Anvari, F., & Tran, H. M. T. (2013). Persona ontology for user centred design professionals. *In Proceedings of the ICIME 4th International Conference on Information Management and Evaluation*.

Aoyama, M. (2005). Persona-and-scenario based requirements engineering for software embedded in digital consumer products. In *Proceedings of the 13th IEEE International Conference on Requirements Engineering.*

Blomquist, A., & Arvola, M. (2002). Personas in action: Ethnography in an interaction design team. In *Proceedings of the 2nd Nordic Conference on Human-Computer Interaction*.

Booth, P. (2008). Rereading fandom: MySpace character personas and narrative identification. *Critical Studies in Media Communication*, *25*(5), 514-536.

Brooks, J. L. (2012). Counterbalancing for serial order carryover effects in experimental condition orders. *Psychological Methods*, *17*(4), 600-614.

Chapman, C., & Milham, R. P. (2006). The personas' new clothes: Methodological and practical arguments against a popular method. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(5), 634-636.

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Hate is not binary: Studying abusive behavior of #GamerGate on Twitter. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*.

Chowdhury, S. A., Mubarak, H., Abdelali, A., Jung, S., Jansen, B. J., & Salminen, J. (2020). A multi-platform Arabic news comment dataset for offensive language detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*.

Cooper, A. (1999). *The inmates are running the asylum: Why high tech products drive us crazy and how to restore the sanity*. New York, NY: Pearson.

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated hate speech detection and the problem of offensive language. In *Proceedings of 11th International AAAI Conference on Web and Social Media*.

Del Vicario, M., Vivaldo, G., Bessi, A., Zollo, F., Scala, A., Caldarelli, G., & Quattrociocchi, W. (2016). Echo Chambers: Emotional Contagion and Group Polarization on Facebook. *Scientific Reports*, *6.*

Eslami, M., Krishna Kumaran, S. R., Sandvig, C., & Karahalios, K. (2018). Communicating algorithmic process in online behavioral advertising. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Faily, S., & Flechais, I. (2011). Persona cases: A technique for grounding personas. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Fiesler, C., & Proferes, N. (2018). “Participant” perceptions of Twitter research ethics. *Social Media+ Society*, *4*(1).

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research*, *20*(2), 303-315.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39-50.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, *51*(4), 1-30.

Fortuna, P., Soler-Company, J., & Wanner, L. (2021). How well do hate speech, toxicity, abusive and offensive language classification models generalize across datasets? *Information Processing & Management*, *58*(3).

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*.

Goodman-Deane, J., Waller, S., Demin, D., González-de-Heredia, A., Bradley, M., Clarkson, J. P., & Clarkson, J. P. (2018). Evaluating inclusivity using quantitative personas. *In the Proceedings of Design Research Society Conference.*

Grudin, J. (2006). Why personas work: The psychological evidence. In J. Pruitt & T. Adlin (Eds.), *The persona lifecycle* (pp. 642-663). New York, NY: Morgan Kaufmann.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2009). *Multivariate data analysis* (7th ed.). New York, NY: Pearson.

Harris, K. A., Spengler, P. M., & Gollery, T. J. (2016). Clinical judgment faith bias: Unexpected findings for psychology research and practice. *Professional Psychology: Research and Practice*, *47*(6), 391-401.

Hill, C. G., Haag, M., Oleson, A., Mendez, C., Marsden, N., Sarma, A., & Burnett, M. (2017). Gender-inclusiveness personas vs. stereotyping: Can we have it both ways? In *Proceedings of the CHI Conference on Human Factors in Computing Systems.*

Hoewe, J. (2017). Manipulation check. In J. Matthes, R. Potter, & C. S. Davis (Eds.), *The international encyclopedia of communication research methods*. Oxford, UK: Wiley-Blackwell.

Hoffmann, A. L., & Jonas, A. (2016). Recasting justice for Internet and online industry research ethics. In M. Zimmer & K. Kinder-Kuranda (Eds.), *Internet research ethics for the social age: New cases and challenges.* Bern, Switzerland: Peter Lang.

Holmgard, C., Green, M. C., Liapis, A., & Togelius, J. (2018). Automated playtesting with procedural personas with evolved heuristics. *IEEE Transactions on Games*, *11*(4), 352-362.

Jansen, A., Van Mechelen, M., & Slegers, K. (2017). Personas and behavioral theories: A case study using self-determination theory to construct overweight personas. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Jansen, B. J., Jung, S., Chowdhury, S. A., & Salminen, J. (2021a). Persona analytics: Analyzing the stability of online segments and content interests over time using non-negative matrix factorization. *Expert Systems with Applications*, *185*.

Jansen, B., Salminen, J., Jung, S., & Guan, K. (2021b). *Data-driven personas*. *Synthesis Lectures on Human-Centered Informatics*, *14*(1), i-317

Jung, S.-G., Salminen, J., An, J., Kwak, H., & Jansen, B. J. (2018a). Automatically conceptualizing social media analytics data via personas. In *Proceedings of the International AAAI Conference on Web and Social Media.*

Jung, S.-G., Salminen, J., & Jansen, B. J. (2019). Personas changing over time: Analyzing variations of data-driven personas during a two-year period. *In Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Jung, S.-G., Salminen, J., Kwak, H., An, J., & Jansen, B. J. (2018b). Automatic persona generation (APG): A rationale and demonstration. In *Proceedings of the Conference on Human Information Interaction & Retrieval*.

Kanekar, S., Bulsara, R. M., Duarte, N. T., & Kolsawalla, M. B. (1981). Perception of an aggressor and his victim as a function of friendship and retaliation. *The Journal of Social Psychology*, *113*(2), 241-246.

Kittur, A., Chi, E. H., & Suh, B. (2009). What's in Wikipedia? Mapping topics and conflict using socially annotated category structure. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Kocoń, J., Figas, A., Gruza, M., Puchalska, D., Kajdanowicz, T., & Kazienko, P. (2021). Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, *58*(5).

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788-791.

LeRouge, C., Ma, J., Sneha, S., & Tolle, K. (2013). User profiles and personas in the design and development of consumer health technologies. *International Journal of Medical Informatics*, *82*(11), e251-268.

Long, F. (2009). Real or imaginary: The effectiveness of using personas in product design. In *Proceedings of the Irish Ergonomics Society Annual Conference*.

Ma, J., & LeRouge, C. (2007). Introducing user profiles and personas into information systems development. In *Proceedings of the Americas Conference on Information Systems.*

Marsden, N., & Haag, M. (2016). Stereotypes and politics: Reflections on personas. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Marsden, N., Hermann, J., & Pröbster, M. (2017a). Developing personas, considering gender: A case study. In *Proceedings of the 29th Australian Conference on Computer-Human Interaction.*

Marsden, N., & Pröbster, M. (2019). Personas and identity: Looking at multiple identities to inform the construction of personas. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.*

Marsden, N., Pröbster, M., Haque, M. E., & Hermann, J. (2017b). Cognitive styles and personas: Designing for users who are different from me. In *Proceedings of the 29th Australian Conference on Computer-Human Interaction*.

Matthews, T., Judge, T., & Whittaker, S. (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*.

McGinn, J. J., & Kotamraju, N. (2008). Data-driven persona development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Mejova, Y., Zhang, A. X., Diakopoulos, N., & Castillo, C. (2014). Controversy and sentiment in online news. *arXiv*. Retrieved from http://arxiv.org/abs/1409.8152

Miaskiewicz, T., & Kozar, K. A. (2011). Personas and user-centered design: How can personas benefit product design processes? *Design Studies*, *32*(5), 417-430.

Mijač, T., Jadrić, M., & Ćukušić, M. (2018). The potential and issues in data-driven development of web personas. In *Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics.*

Minichiello, A., Hood, J. R., & Harkness, D. S. (2018). Bringing user experience design to bear on STEM education: A narrative literature review. *Journal for STEM Education Research*, *1*(1-2), 7-33.

Molenaar, L. (2017). *Data-driven personas: Generating consumer insights with the use of clustering analysis from big data* (master's thesis). Delft University of Technology, Delft, Netherlands.

Morris, M. W., Leung, K., & Iyengar, S. S. (2004). Person perception in the heat of conflict: Negative trait attributions affect procedural preferences and account for situational and cultural differences. *Asian Journal of Social Psychology*, *7*(2), 127-147.

Nielsen, L. (2004). *Engaging personas and narrative scenarios* (doctoral thesis). Copenhagen Business School, Copenhagen, Denmark.

Nielsen, L. (2019). *Personas—user focused design* (2nd ed.). Berlin: Springer.

Nielsen, L. (2002). From user to character: An investigation into user-descriptions in scenarios. In *Proceedings of the 4th Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*.

Nielsen, L., Hansen, K. S., Stage, J., & Billestrup, J. (2015). A template for design personas: Analysis of 47 persona descriptions from Danish industries and organizations. *International Journal of Sociotechnology and Knowledgement Development*, *7*(1), 45-61.

Nielsen, L., Jung, S.-G., An, J., Salminen, J., Kwak, H., & Jansen, B. J. (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*.

Nielsen, L., Nielsen, K. S., Stage, J., & Billestrup, J. (2013). Going global with personas. In P. Kotzé, G. Marsden, G. Lindgaard, J. Wesson, & M. Winckler (Eds.), *Human-computer interaction* (LNCS vol. 8120). Berlin: Springer.

Nielsen, L., & Hansen, K. S. (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*.

Nieters, J. E., Ivaturi, S., & Ahmed, I. (2007). Making personas memorable. *In Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Nisbett, R. E., & Wilson, T. D. (1977). The halo effect: Evidence for unconscious alteration of judgments. *Journal of Personality and Social Psychology*, *35*(4), 250.

Palan, S., & Schitter, C. (2018). Prolific.ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22-27.

Pamungkas, E. W., Basile, V., & Patti, V. (2021). A joint learning approach with knowledge injection for zero-shot cross-lingual hate speech detection. *Information Processing & Management*, *58*(4).

Parveen, F., Jaafar, N. I., & Ainin, S. (2015). Social media usage and organizational performance: Reflections of Malaysian social media managers. *Telematics and Informatics*, *32*(1), 67-78.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology*, *70*, 153-163.

Pröbster, M., Hermann, J., & Marsden, N. (2019). Personas and persons—an empirical study on stereotyping of personas. In *Proceedings of Mensch Und Computer.*

Pruitt, J., & Grudin, J. (2003). Personas: Practice and theory. In *Proceedings of the Conference on Designing for User Experiences*.

Psychology Research and Reference. (2018). *Person perception*. Retrieved from https://psychology.iresearchnet.com/social-psychology/social-cognition/person-perception/

Putnam, C., Kolko, B., & Wood, S. (2012). Communicating about users in ICTD: Leveraging HCI personas. In *Proceedings of the 5th International Conference on Information and Communication Technologies and Development*.

Rentsch, J. R., & Woehr, D. J. (2004). Quantifying congruence in cognition: Social relations modeling and team member schema similarity. In E. Salas & S. M. Fiore (Eds.), *Team cognition: Understanding the factors that drive process and performance* (pp. 11-31). Worcester, MA: American Psychological Association

Reysen, S. (2005). Construction of a new scale: The Reysen likability scale. *Social Behavior and Personality*, *33*(2), 201-208.

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.*

Rönkkö, K., Hellman, M., Kilander, B., & Dittrich, Y. (2004). Personas is not applicable: Local remedies interpreted in a wider context. In *Proceedings of the 8th Conference on Participatory Design: Artful Integration: Interweaving Media, Materials and Practices.*

Salminen, J., Almerekhi, H., Milenković, M., Jung, S., An, J., Kwak, H., & Jansen, B. J. (2018a). Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Proceedings of the International AAAI Conference on Web and Social Media.*

Salminen, J., Froneman, W., Jung, S., Chowdhury, S., & Jansen, B. J. (2020a). The ethics of data-driven personas. *In Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Salminen, J., Guan, K., Jung, S., Chowdhury, S. A., & Jansen, B. J. (2020b). A literature review of quantitative persona creation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Salminen, J., Jansen, B. J., An, J., Kwak, H., & Jung, S. (2018b). Are personas done? Evaluating their usefulness in the age of digital analytics. *Persona Studies*, *4*(2), 47-65.

Salminen, J., Jansen, B. J., An, J., Kwak, H., & Jung, S.-G. (2019a). Automatic persona generation for online content creators: Conceptual rationale and a research agenda. In L. Nielsen (Ed.), *Personas—user focused design* (pp. 135-160). Berlin: Springer.

Salminen, J., Jung, S., An, J., Kwak, H., Nielsen, L., & Jansen, B. J. (2019b). Confusion and information triggered by photos in persona profiles. *International Journal of Human-Computer Studies*, *129*, 1-14.

Salminen, J., Jung, S., Chowdhury, S. A., Sengün, S., & Jansen, B. J. (2020c). Personas and analytics: A comparative user study of efficiency and effectiveness for a user identification task. Proceedings of the CHI Conference of Human Factors in Computing Systems.

Salminen, J., Jung, S., & Jansen, B. J. (2019c). The future of data-driven personas: A marriage of online analytics numbers and human attributes. In *Proceedings of the 21st International Conference on Enterprise Information Systems*.

Salminen, J., Jung, S., Santos, J. M., & Jansen, B. J. (2019d). 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*, *36*(6), 568-590.

Salminen, J., Jung, S., Santos, J. M., Kamel, A. M., & Jansen, B. J. (2021). Picturing it! The effect of image styles on user perceptions of personas. In *the Proceedings of ACM Human Factors in Computing Systems (CHI'21)*.

Salminen, J., Jung, S.-G., An, J., Kwak, H., & Jansen, B. J. (2018c). Findings of a user study of automatically generated personas. *In Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Salminen, J., Nielsen, L., Jung, S.-G., An, J., Kwak, H., & Jansen, B. J. (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.*

Salminen, J., Santos, J. M., Jung, S., Eslami, M., & Jansen, B. J. (2019e). Persona transparency: Analyzing the impact of explanations on perceptions of data-driven personas. *International Journal of Human-Computer Interaction*, *36*(8), 788-800.

Salminen, J., Santos, J. M., Kwak, H., An, J., Jung, S., & Jansen, B. J. (2020d). Persona perception scale: Development and exploratory validation of an instrument for evaluating individuals' perceptions of personas. *International Journal of Human-Computer Studies*, *141*.

Salminen, J., Sengün, S., Corporan, J., Jung, S., & Jansen, B. J. (2020e). Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. *PLOS ONE*, *15*(2), e0228723.

Salminen, J., Şengün, S., Kwak, H., Jansen, B. J., An, J., Jung, S., Vieweg, S., & Harrell, F. (2018e). From 2,772 segments to five personas: Summarizing a diverse online audience by generating culturally adapted personas. *First Monday*, *23*(6).

Salminen, J., Veronesi, F., Almerekhi, H., Jung, S., & Jansen, B. J. (2018f). Online hate interpretation varies by country, but more by individual: A statistical analysis using crowdsourced ratings. In *Proceedings of the 5th International Conference on Social Networks Analysis, Management and Security.*

Sengün, S., Salminen, J., Mawhorter, P., Jung, S., & Jansen, B. J. (2019). Exploring the relationship between game content and culture-based toxicity: A case study of League of Legends and MENA players. In *the Proceedings of the 30th ACM Conference on Hypertext and Social Media.*

Singer, T., & Klimecki, O. M. (2014). Empathy and compassion. *Current Biology*, *24*(18), R875-R878.

Spiliotopoulos, D., Margaris, D., & Vassilakis, C. (2020). Data-assisted persona construction using social media data. *Big Data and Cognitive Computing*, *4*(3), 1-14.

Stevenson, P. D., & Mattson, C. A. (2019). The personification of big data. In *Proceedings of the Design Society: International Conference on Engineering Design*, *1*(1), 4019-4028.

Türkay, S., Formosa, J., Adinolf, S., Cuthbert, R., & Altizer, R. (2020). See no evil, hear no evil, speak no evil: How collegiate players define, experience and cope with toxicity. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Ullmann, S., & Tomalin, M. (2019). Quarantining online hate speech: Technical and ethical perspectives. *Ethics and Information Technology*, *22*, 69-80.

Wispé, L. (1986). The distinction between sympathy and empathy: To call forth a concept, a word is needed. *Journal of Personality and Social Psychology*, *50*(2), 314-321.

Zhang, X., Brown, H.-F., & Shankar, A. (2016). Data-driven personas: Constructing archetypal users with clickstreams and user telemetry. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Appendix

Ethical questions about using social media data have become more important for researchers in computational social science and HCI (Ullmann & Tomalin, 2019). According to the calls to make ethical choices more explicit in studies in these fields (Fiesler & Proferes, 2018; Hoffmann & Jonas, 2016), we state the following ethical choices that we made in this study:

- We generated the personas using aggregated, non-personally identifiable information
- We adhered to the data-collection platform's terms of service
- We informed the recruited participants on the content's explicit nature and acquired their consent for participation

With these choices, we focused on mitigating potential harm to individuals' privacy and the psychological harm that the toxic comments imposed on them. Due to the study's nature, we could not purposefully block the explicit content in the toxic comments.

About the Authors

Joni Salminen works as a scientist in the Qatar Computing Research Institute and is also affiliated with the Turku School of Economics at the University of Turku. His research interests relate to personas, human-computer interaction, online hate, user and customer segmentation, and social media platforms.

Soon-gyo Jung is a software engineer focused on data-driven/data-intensive systems in the social computing group at Qatar Computing Research Institute (QCRI), Doha, Qatar. He is currently strongly interested in computational social science, especially exploring its significant impact on society and how people communicate and share their culture with others. He received a BE degree in computer software from the Kwangwoon University, Seoul, Korea, in 2014, and a MS degree in electrical and computer engineering from the Sungkyunkwan University, Suwon, Korea, in 2016.

João M. Santos is a researcher at Instituto Universitário de Lisboa (ISCTE-IUL) specializing in data analysis and science & technology studies. His research interests are centered around the social sciences, in general, with particular focus on higher education, public policy, data analysis, and psychology.

Bernard J. Jansen is a Principal Scientist in the social computing group of the Qatar Computing Research Institute. He is a graduate of West Point and has a PhD in computer science from Texas A&M University. He is Editor-in-Chief of *Information Processing & Management* (Elsevier).

Transactions on Human – Computer Interaction