Intentionally Biasing User Representation?: Investigating the Pros and Cons of Removing Toxic Quotes from Social Media Personas

Joni Salminen
University of Vaasa, Vaasa, Finland
jonisalm@uwasa.fi

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

Bernard J. Jansen
Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar
bjansen@hbku.edu.qa

ABSTRACT

Algorithmically generated personas can help organizations understand their social media audiences. However, when using algorithms to create personas from social media user data, the resulting personas may contain toxic quotes that negatively affect content creators' perceptions of the personas. To address this issue, we have implemented toxicity detection in an algorithmic persona generation system capable of using tens of millions of social media interactions and user comments for persona creation. On the system's user interface, we provide a feature for content creators using the personas to turn on or off toxic quotes, depending on their preferences. To investigate the feasibility of this feature, we conducted a study with 50 professionals in the online publishing domain. The results show varied reactions, including hate-filter critics, hate-filter advocates, and those in between. Although personal preferences play a role, the usefulness of toxicity filtering appears primarily driven by the work task – specifically the type and topic of stories the content creator seeks to create. We identify six use cases where a toxicity filter is beneficial. For system development, the results imply that it is beneficial to give content creators the option to view or not view toxic comments, rather than making this decision in their stead. We also discuss the ethical implications of removing toxic quotes in algorithmically generated personas, including potentially biasing the user representation.

CCS CONCEPTS

• **Human-centered computing** → Human computer interaction (HCI).

KEYWORDS

Algorithmically Generated Personas, Social Media Personas, Online Toxicity, Machine Learning, Feasibility Study

ACM Reference Format:
Joni Salminen, Soon-Gyo Jung, and Bernard J. Jansen. 2022. Intentionally Biasing User Representation?: Investigating the Pros and Cons of Removing Toxic Quotes from Social Media Personas. In *Nordic Human-Computer Interaction Conference (NordiCHI '22), October 08–12, 2022, Aarhus, Denmark.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3546155.3546647

1 INTRODUCTION

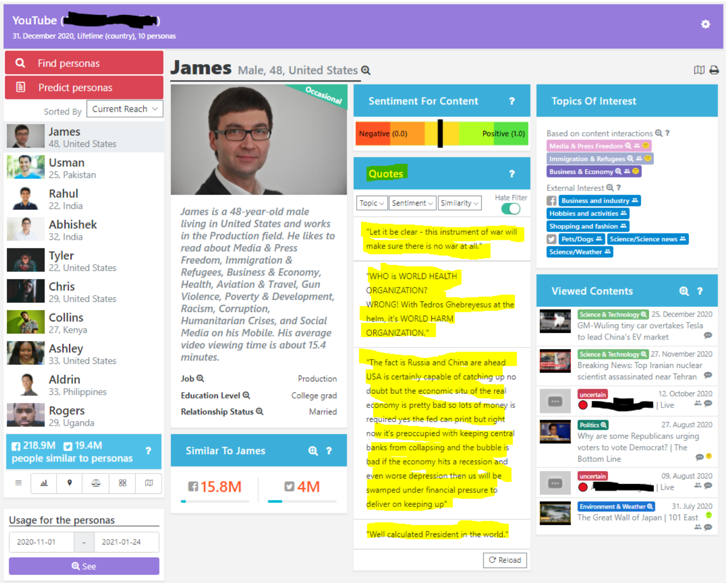
Personas are fictitious human beings that represent users of a system, product, or online content [16, 84]. Personas support user-centered activities [8] in human-computer interaction (HCI) and related fields, as they increase the empathetic understanding of users [79]. Because personas aggregate users under specific types, designers, developers and content creators can consider the persona's needs and wants in their work [25, 57], and communicate about users to others in the organization [10, 49, 53] in an easily approachable way – similar to a 'real person'. One prominent area of persona application is online content creation—e.g., digital journalism, research publication services, professional bloggers, social media influencers, etc.—where personas are used for understanding digital audiences and online news readers [15, 29, 32, 75].

Even though personas are well-established in industry and research [28, 50], manual persona creation has drawbacks, including small sample sizes, subjective analysis methods, time and cost factors, and the difficulty to account for changing user behavior over time [13, 33, 63, 64]. To address these drawbacks and to keep up with the ever-increasing volume of quantitative social media user data, researchers have developed methodologies and systems for algorithmically generated personas (AGPs) that are created from online analytics and social media user data [4, 5, 45, 48, 75, 88]. AGPs are enabled by three major data-driven trends [1, 61]: (1) *online platforms* that provide aggregated statistics about users [81], (2) *data science algorithms* that help automate persona creation analyses [86], and (3) *online interaction techniques* that enable persona users' interaction with the personas via responsive systems and user interfaces (UI) [30].

AGPs typically display user information in persona profiles (see Figure 1) which also contain quotes (i.e., comments by users in online conversations) that are intended to reflect the personality and attitudes of the persona [51]. When creating personas for online content creators to represent social media user data, any toxic quotes may affect online content creators' impressions of what kind of person the persona is [65, 66, 71]. Following the creation of AGPs from social media user data, researchers have observed an increasing likelihood of *toxic comments appearing on algorithmically generated personas* [65, 71]. A toxic comment is defined as commenting involving harmful intent, often targeting an individual or group [59]. These toxic comments may automatically be selected

NordiCHI '22, October 08–12, 2022, Aarhus, Denmark

ACM ISBN 978-1-4503-9699-8/22/10…$15.00
https://doi.org/10.1145/3546155.3546647



YouTube ([redacted])
31. December 2020, Lifetime (country), 10 personas

Find personas
Predict personas
Sorted By: Current Reach

James, 48, United States
Usman, 25, Pakistan
Rahul, 22, India
Abhishek, 32, India
Tyler, 22, United States
Chris, 29, United States
Collins, 27, Kenya
Ashley, 33, United States
Aldrin, 33, Philippines
Rogers, 29, Uganda

218.9M 19.4M people similar to personas

Usage for the personas
2020-11-01 - 2021-01-24
See

James Male, 48, United States

Occasional

James is a 48-year-old male living in United States and works in the Production field. He likes to read about Media & Press Freedom, Immigration & Refugees, Business & Economy, Health, Aviation & Travel, Gun Violence, Poverty & Development, Racism, Corruption, Humanitarian Crises, and Social Media on his Mobile. His average video viewing time is about 15.4 minutes.

Job: Production
Education Level: College grad
Relationship Status: Married

Similar To James
15.8M 4M

Sentiment For Content
Negative (0.0) Positive (1.0)

Quotes
Topic Sentiment Similarity Hate Filter

"Let it be clear - this instrument of war will make sure there is no war at all."

"WHO is WORLD HEALTH ORGANIZATION?
WRONG! With Tedros Ghebreyesus at the helm, it's WORLD HARM ORGANIZATION."

"The fact is Russia and China are ahead USA is certainly capable of catching up no doubt but the economic situ of the real economy is pretty bad so lots of money is required yes the fed can print but right now it's preoccupied with keeping central banks from collapsing and the bubble is bad if the economy hits a recession and even worse depression then us will be swamped under financial pressure to deliver on keeping up"

"Well calculated President in the world."

Reload

Topics Of Interest
Based on content interactions
Media & Press Freedom
Immigration & Refugees
Business & Economy
External Interest
Business and industry
Hobbies and activities
Shopping and fashion
Pets/Dogs
Science/Science news
Science/Weather

Viewed Contents
Science & Technology | 25. December 2020 | GM-Wuling tiny car overtakes Tesla to lead China's EV market
Science & Technology | 27. November 2020 | Breaking News: Top Iranian nuclear scientist assassinated near Tehran
uncertain | 12. October 2020 | [redacted] | Live
Politics | 27. August 2020 | Why are some Republicans urging voters to vote Democrat? | The Bottom Line
uncertain | 09. August 2020 | [redacted] | Live
Environment & Weather | 31. July 2020 | The Great Wall of Japan | 101 East

Figure 1: An automatically created social media persona from an online news channel's data, served via an interactive persona system. Persona profiles typically contain quotes reflecting the persona's attitudes and opinions (see yellow highlights in the middle of the figure). The organization's name is masked.

for certain AGPs by the algorithm from social media user data [47]. Therefore, depending on their prevalence in the user data, toxic comments may appear in automatically generated personas without the persona creators deliberately selecting them for the persona profile. These toxic quotes may have multiple negative effects on content creators' perceptions of the AGPs:

- **The impressions of the persona may become contaminated**. When algorithms select toxic quotes, the toxicity in these comments is interpreted by content creators to reflect the attitudes of the persona, even when such toxicity would not represent the majority views of the users that the persona represents [68].
- The toxicity may reduce the content creator's empathy toward the persona when creating content for the segment that the persona represents. Instead of understanding the persona for the content creation task, the content creator might be resentful and thus reluctant to design for a toxic person. Such a loss of empathy would signify losing the primary benefit of deploying personas in the first place [21, 50, 68].
- **Toxic quotes may attract content creators' attention to irrelevant information.** Instead of learning about the persona's needs, content creators may be compelled to focus on one or several negative comments that do not necessarily provide any valuable information for decision-making [3] on content creation and other user-centric tasks. Thereby, toxic quotes can distract the persona user from their professional task.

A particular concern is that even a single toxic quote may negatively affect the persona user's perception of the persona, as persona users tend to be draw major inferences even from small details in the persona profile [65]. In social psychology, this is referred to as the reputation spoiling effect [54]. This effect amplifies the issue because, even though toxicity does take place in social media, *most* online users are *not* toxic [62]. Therefore, the impact of one or several toxic quotes may be unjustifiably strong in terms of user perceptions and the accurate portrayal of user segments.

All of these aspects are also connected with the theme of ethics in computing systems, and intelligent and autonomous systems in particular [23, 60]. Normatively speaking, when creating personas automatically, the humans applying the algorithms should consider the ethical dimensions of the process, instead of transferring control to the autonomous system [40]. As machine learning (ML) and artificial intelligence (AI) technologies progress, their implementation is no longer hindered by technical abilities, but more by human factors. Consequently, this raises a question as to under what circumstances should automation be used to affect user representation, such as persona creation, user modeling, and user segmentation? There is a scarcity of research addressing this issue, and our study focuses on the context of AGPs, and discusses a specific case within the theme of merging AI with HCI.

We investigate the *toxic quote problem in data-driven personas* by implementing an ML model for toxicity detection that scores each comment before constructing the persona profile. Then, we provide a feature in a persona system's UI to turn off toxic quotes. Content creators using the personas can thus choose to include or exclude toxic comments from the personas they are using, giving them control over the information they see. We also conduct a feasibility study among content creator professionals to investigate their perceptions of this filtering feature. With this research, we provide crucial findings on perceptions of human-AI collaboration [36] in the context of online content creation, where AGPs may inherit undesirable aspects due to the application of algorithms for social media personas. The main contribution is an overview of the arguments in favor and against the use of this tool for content creation. We present a detailed analysis of these arguments and provide reasoning for why the filtering of toxic quotes in personas could be useful, and also why it might be harmful.

2 RELATED LITERATURE

The Nordic tradition has focused on 'design personas' [10, 11, 53, 77], intended to assist product/UI/UX designers and software developers. This concept is rooted in Cooper's 'user personas' [16] that are representations of current or potential software users. Online content creation can be seen as a special case of design tasks, in which the designer (content creator) designs (creates) content that is targeted or tailored to specific audience segments (personas).

Research has shown that persona users' perceptions of the personas they are exposed to are vital for deploying personas in real organizations. Critical perceptions mentioned in the literature include accuracy [13], credibility [43], immersion and empathy [50], perceived personality [6, 7], trust [10], and the persona's usefulness [16, 63]. The (in)consistency of the information that stakeholders see in the persona profiles contribute to their persona perceptions [11].

Research has also established that perceptions of personas differ individually, and involve stereotypical thinking or subjective interpretation due to the unique backgrounds and experiences of persona users [27, 41, 71]. Therefore, when assessing users' perceptions of persona, it is logical to adopt the concept of *person* perception, i.e., the "general tendency to form impressions of other people" [58] (p. 1) from social psychology [25]. These beliefs that people attribute to others can relate to appearance, demographic attributes, behaviors, and other human characteristics [2]. As such, persona perceptions are human aspects of the personas that content creators consider when forming their impression of the personas, based on available information [25, 49]. Moreover, the above premise denotes a departure from the idea that personas should be solely evaluated by measuring how *accurate* they are in a technical sense [13], and additionally, psychological and demographic factors also play an important role [56]. While researchers should attempt to create accurate personas that represent the data faithfully, it is also vital to create personas that are perceived productively [10], i.e., to support the users' (e.g., content creators') goals.

Stakeholders' perceptions of personas are *especially* meaningful when algorithms are used for designing AGPs [20], which involves making decisions about what information to include in the AGPs. When the quotes of AGPs are selected automatically from the content the persona has most engaged with [4], the personas' degree of toxicity correlates with the toxicity of the comments, a situation which is aggravated by the existence of bot and troll accounts on social media [9, 80]. Thus, the more toxic comments there are, the more likely the persona will appear to be toxic to content creators. As laid out above, the inclusion of toxic comments in AGPs can have adverse side effects [52, 66, 71]. It is possible that any number of toxic comments may make the whole persona appear toxic for a content creator, and they may not be able to relate to a toxic persona (or be less able to do so), thus eradicating the empathy advantages associated with personas. Toxic quotes can also direct attention away from persona information that is more important for the content creation task, such as the persona's goals, overall sentiment, or topics of interest. Particularly, Salminen et al. [68] carried out a controlled user experiment showing 496 participants both toxic and non-toxic versions of the same manually created personas, and then asked about the participants' impressions. The results indicated that participants found the non-toxic personas (i.e., those that had no toxic quotes) to be more empathetic, likable, and credible. The participants also identified more strongly with the non-toxic personas, and were more willing to use them for design tasks. These findings imply that a toxicity of AGPs results in downstream effects for the perception and use of personas. However, it is unclear how the toxicity problem may be addressed in algorithmically generated personas, which forms a major research gap. To address this research gap, the current work implements a filtering module for toxic comments in an algorithmic persona generation system, and evaluates the module by gathering feedback from professionals in the field.

3 EXPLORATORY ANALYSIS OF TOXICITY IN AUTOMATICALLY GENERATED PERSONAS

3.1 Overview

One of the most prominent domains exposed to online toxicity is online content creation, and especially news content. Due to their reporting of political, religious, and other controversial topics [37, 59, 74], news and media organizations are typical targets of toxicity. AGPs generated from news audience data are a prime example of the toxic quote problem in persona generation. In order to validate this idea, in this research, we compared the personas created from the data of two organizations—a news channel and a non-profit—to investigate how prevalent the toxic quotes were when using algorithmic persona generation. News channels are routinely attacked by toxic commentators [59], and the attacks mainly target the channels themselves, religious groups, and nationalities. Because of the high prevalence of toxic comments, these toxic comments are often included as 'quotes' for the personas, and as a consequence, the AGPs themselves may appear as repulsive and toxic.

3.2 Data Collection and Persona Generation

The study dataset was collected using the YouTube Analytics API[1], in accordance with Google's terms of service and a research agreement with a major content creation organization that permits the use of the data for research. The data contains no personally identifiable information concerning individual users, apart from the username used in the comments (which are available on the YouTube website and are publicly viewable). We do not show these usernames in the generated personas; instead, the personas' names are selected based on a probabilistic algorithm [34]. In this research, we generated social media audience personas using an algorithmic persona generation system that is state-of-the-art in persona creation [4, 5]. Figure 1 shows an example of a generated persona from this system. Here, we briefly explain the system's procedure for

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

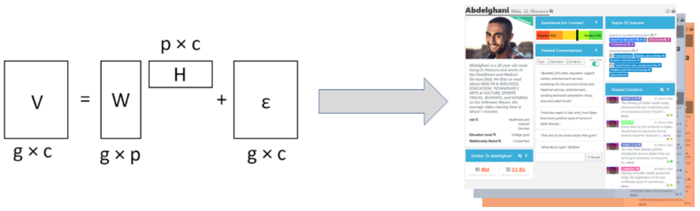


Figure 2: Algorithmic persona generation. Non-negative matrix factorization derives latent patterns from user data (left-hand side of the illustration). The algorithmic persona generation system then enriches these latent patterns with personified information such as name, picture, demographics, and so on (right-hand side of the illustration). See technical details in [5].

generating AGPs (Figure 2). However, a more detailed technical description of the system can be found in previously published work [32, 35] that explains the functionalities and algorithmic processes underlying the system [4, 5].

The algorithmic persona creation system applies the following steps to generate personas from social media user data:

- **Step 1:** The persona system creates an interaction matrix that has online content pieces (e.g., videos) as columns, demographic user groups as rows, and the engagement counts of each group for each content as elements.
- **Step 2:** The persona system applies the non-negative matrix factorization (NMF) algorithm [38] to this interaction matrix to detect a fixed number of latent engagement behaviors.
- **Step 3:** The persona system chooses representative demographic traits for each engagement behavior from the highest weights provided by the NMF algorithm.
- **Step 4:** The persona system finalizes the personas by adding more information in the representative demographic groups for each persona, such as name, picture, sociographic, and topics of interest.

These personas are provided to users (i.e., content creators) via an algorithmic persona generation system, as shown in Figure 1. By default, each AGP displays four comments. The comments in these personas are algorithmically picked at random from the content the persona has most interacted with. Neither of the organizations actively moderates their content, so the chosen comments represent unfiltered user-generated comments. These comments can be refreshed using a feature in the system (see the "Reload" button in Figure 1, under the comments).

3.3 Evaluating Toxicity in Algorithmically Generated Personas

We first conducted an exploratory experiment to judge the prevalence of toxicity comments in the social media channel of a news and media organization, relative to a non-news and media organization (in this case, a non-profit organization). We chose three personas generated automatically for each organization and refreshed the comments widget eight times, amounting to 32 total comments that were used to evaluate each persona. We evaluated three personas for each organization, as the personas' content may vary by toxicity. In other words, there were 3 × 32 = 96 comments per organization, and 192 in total. Two researchers manually coded the comments for toxic content (**1** = Has toxic content, **0** = Does not have toxic content). The inter-rater agreement was calculated using Cohen's Kappa, which indicated a near-perfect agreement (k = 0.90). Any disagreed cases (n = 5, 2.6%) were discussed case-by-case by the two researchers, and a final label for each was determined in unison.

The results in Table 1 show that the news and media channel was subject to a considerably larger extent of toxic commenting, as we expected. The difference in toxicity prevalence is statistically significant, with 22.9% of the personas' comments for the news and media organization and 10.4% for the non-profit labeled as toxic [X^2 (1, N = 192) = 5.4, p = .020]. The result underlines not only the fact that the "natural" toxicity of social media commenting varies by context (channel), but also that this toxicity is effectively inherited by the resulting AGPs, as we suggested in the introduction. In addition, as Figure 3 portrays, there are some differences in the toxicity of the AGPs, arising from the fact that different content (videos, in this case) garners different levels of toxicity, which has been described in previous research as "topic-driven toxicity" [74]. As such, there is a high chance of toxic quotes emerging in the personas of online news organizations when using algorithmically generated persona creation approaches, at least in this particular case. These results provided the motivation for continuing our research.

4 DEVELOPING AN INTERACTIVE SYSTEM FOR FILTERING TOXIC QUOTES

4.1 System Design

Our design philosophy was to provide content creators with the choice of seeing either toxic or non-toxic personas by disabling or enabling toxic quotes from the persona system. For this purpose, we developed an ML-based module called *PURIFIER*. Using this module, the persona generation algorithms automatically derive the quotes (i.e., user comments) in the AGPs from the content of the organization's social media channel whose personas the system creates.

4.2 Model Development

To develop an ML model for toxicity detection, we used four datasets from previous peer-reviewed research (see [62]). We chose these four datasets to improve applicability across multiple social media platforms, as persona systems work with many data sources. Notably, their application has been seen to influence the choice of algorithms and feature representations in previous research [17, 59].

Table 1: Manual toxicity rating (final ratings after researcher discussions).

	News organization	Non-profit organization
Total N (evaluated)	96	96
Toxic	22	10
Toxic ratio	22.9%	10.4%

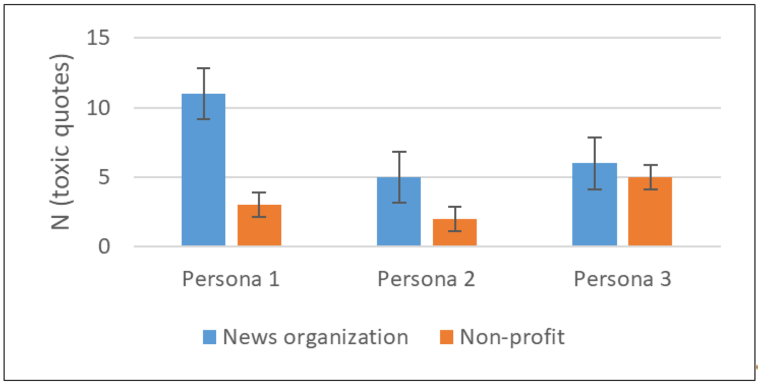


Figure 3: Number of toxic quotes in the evaluated personas. Error bars indicate standard error. All personas have some toxic quotes, but these differ in number. For the news organization, the most toxic persona was the first one shown in the system UI. For the non-profit, the most toxic was the third persona in the system UI. In the system UI, the personas are sorted by their representation of the channel's total audience, with personas with a higher representation being shown first.

Using the combined data that consisted of toxic and non-toxic social media comments, we trained various classification algorithms: Logistic Regression (LR), Naïve Bayes, Support Vector Machines (SVM), Feed-Forward Neural Network (FFNN), and Extreme Gradient Boosting (XGBoost). The features included bag of words (BOW), Term Frequency - Inverted Document Frequency (TF-IDF), GloVe word embeddings [55], and Bidirectional Encoder Representations from Transformers (BERT), i.e., features obtained from a transformer network [18]. The technical details of the ML model development have been reported in related work [62].

It should be clarified that the contribution of the presented work is not the development of a toxicity classifier, but its application and stakeholder evaluation as a tool for filtering toxic content from AI-generated personas. The technical ML evaluation results (see Table 3) show that the XGBoost algorithm with all of the features performed the best (F1 score = 0.924); hence, we chose this model as the basis for the PURIFIER. Using this model, PURIFIER classifies each retrieved social media comment. If the classification probability of the comment to be toxic increases beyond 0.50, the AGP module excludes the comment from being shown in the persona profile if the toxicity filter is enabled.

4.3 Implementing Toxicity Filtering to the Persona System's User Interface

One additional aspect that we must consider when addressing this design challenge is the variety of individual preferences regarding the impact of toxic quotes on their persona perceptions. According to our discussions with the people in the news organization, some individuals find toxic comments offensive and do not want to see them. In contrast, others are genuinely interested in their audience's "dark side". Therefore, the decision to see toxic quotes in AGPs might ultimately be for the persona end user. For this purpose, we implemented a reasonable solution of providing a standard UI element in the form of a "toggle" that filters the toxic quotes out/in, should the user want this option (see Figure 4). We refer to this option as the *Hate Filter*. As well as those who may wish to filter out toxic comments, for example, there may be content creators who *want* to understand the sources and targets of the toxicity [66], so giving users a choice of what they see is crucial for designing information system UIs for content publishing [20].

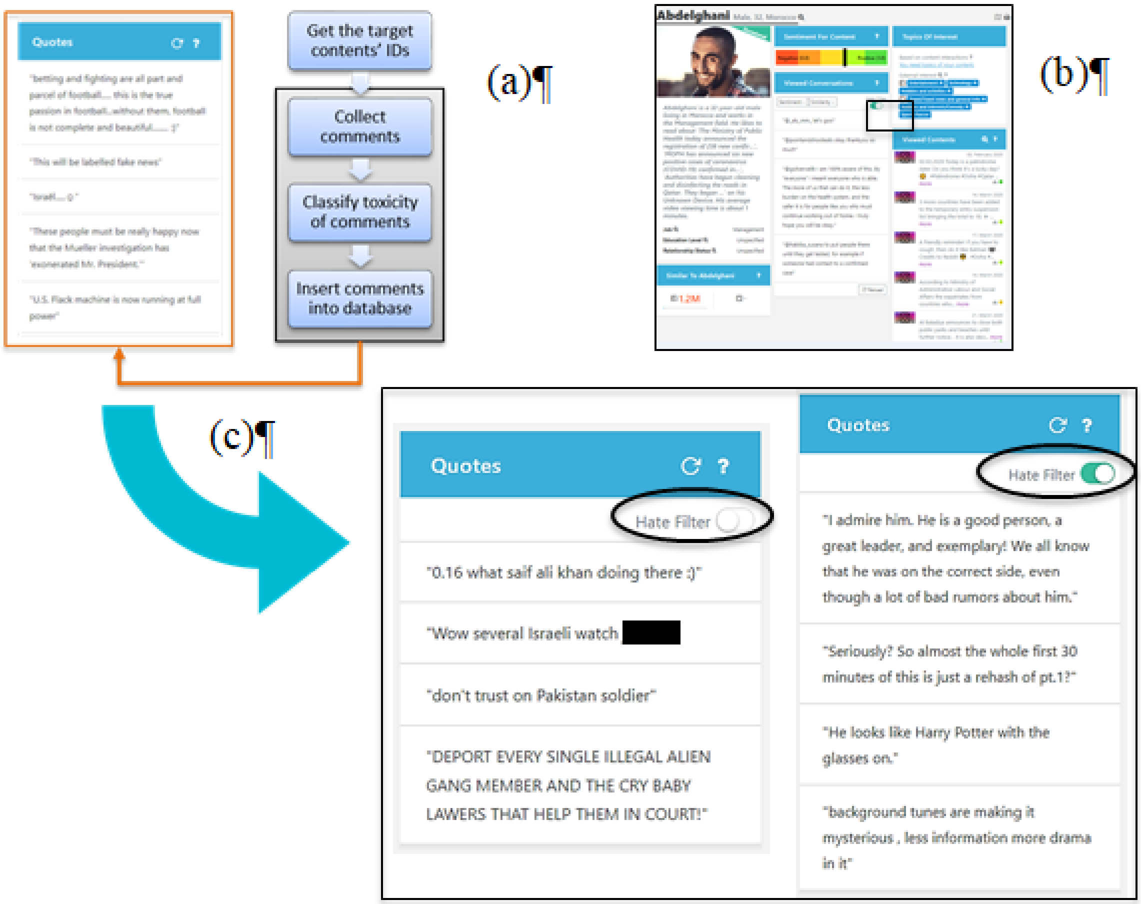
5 FEASIBILITY STUDY

5.1 Purpose

A feasibility study of people working in online content creation was conducted to collect feedback on user perceptions of the toxicity filter feature's usefulness, and the possible risks involved. Measured perceptions and items (see Table 2) were devised to reflect different aspects relevant for the use case of toxicity filtering. For example, usefulness was measured by asking, "In your opinion, how useful is the toxicity filtering feature?" These constructs and items correspond to those deployed in previous quantitative persona user studies (e.g., [69, 72, 73]).

5.2 Participant Recruitment

To recruit participants, we used Prolific, an online survey and study platform that has been deployed in previous research, including


Quotes
Get the target contents' IDs
Collect comments
Classify toxicity of comments
Insert comments into database
(a)¶
Abdelghani
(b)¶
(c)¶
Quotes
Hate Filter
Quotes
Hate Filter

"betting and fighting are all part and parcel of football.... this is the true passion in football..without them, football is not complete and beautiful...... :)"

"This will be labelled fake news"

"Israel..... :) "

"These people must be really happy now that the Mueller investigation has 'exonerated Mr. President.'"

"U.S. Flack machine is now running at full power"

"0.16 what saif ali khan doing there :)"

"Wow several Israeli watch [redacted]

"don't trust on Pakistan soldier"

"DEPORT EVERY SINGLE ILLEGAL ALIEN GANG MEMBER AND THE CRY BABY LAWERS THAT HELP THEM IN COURT!"

"I admire him. He is a good person, a great leader, and exemplary! We all know that he was on the correct side, even though a lot of bad rumors about him."

"Seriously? So almost the whole first 30 minutes of this is just a rehash of pt.1?"

"He looks like Harry Potter with the glasses on."

"background tunes are making it mysterious , less information more drama in it"

Figure 4: Schematics of PURIFIER. (a) Comment retrieval and filtering in the AGP system: identifying the IDs of the persona's most viewed contents, collecting comments from these comments, classifying their toxicity, and storing the comments in a database. (b) The picture shows the Hate Filter implementation in Persona UI. (c) The picture shows the Hate Filter off and on. This option gives the persona users ultimate control over seeing or not seeing toxic quotes in their personas.

Table 2: User feedback measures using a seven-point Likert scale: for C01, 7 = Extremely useful; for C02–07, 7 = Strongly agree.

ID	Construct	Measurement item
C01	Usefulness	In your opinion, how useful is the toxicity filtering feature?
C02	User Experience	Using the toxicity filtering feature would improve my user experience when reviewing persona information.
C03	Insights	Using the toxicity filtering feature would improve my understanding of users.
C04	Realism	Using the toxicity filtering feature would make the personas more realistic.
C05	Risk Factor	I believe that there could be negative consequences when using the toxicity filtering feature.
C06	Effectiveness	I think that the toxicity filtering feature would improve the persona system's effectiveness of portraying user segments.
C07	Trust	I would trust that the toxicity filtering works correctly.

persona user studies [69, 70, 72, 73]. The following sampling criteria were applied to find people working in online content creation or related fields: Age = 23–65 (as we wanted people active in work-life), Industry = Publishing (this was the conceptually closest category available in Prolific), and Student status = No (as we wanted people active in work-life). The platform indicated 220 matching participants who had been active in the past 90 days. Out of these, we recruited 51, which was considered to be an adequate number to provide insights into the matters we focused on. Based on a pre-test with a handful of participants, the answering time for our survey was estimated to be around 10 minutes. Based on this estimate, we set a reward of £1.46 on the Prolific task, which is equivalent to an hourly rate of £8.76. This exceeds the United Kingdom (UK) hourly minimum wage for people over 25 years old, as it stood at the time of the study. Offering a reward level matching at least the minimum wage is considered as fair treatment of online subjects [85], and we chose the UK as the basis because of the panel's large prevalence of UK residents[2].

5.3 Participant Information

We included an attention check question in the survey to control the quality of the responses. Of the 51 obtained responses, one participant failed the attention check and was removed from the study. This left a total number of 50 participants for the analysis. More than half were female (n=31, 62.0%), the rest male (n=19, 38.0%). Their average age was 36.2 years (SD=10.0). Slightly over one-third (36.0%) of the participants were not familiar with personas before this study. Close to two-thirds (64.0%) knew of personas, with 44.0% not having used them professionally. Twenty percent (20.0%) knew of personas and had employed them at least once. This distribution is similar to that observed in other persona user studies [67, 70, 71]. To ensure an equitable baseline, all participants were introduced to the concept of personas and their usage. The participants' average work experience in content creation or the publishing industry was 8.28 years (SD=7.28). The participants represented 12 nationalities. Most of the participants (56.0%) were from the UK, with the rest of the participants coming from different parts of the world, e.g., the United States (n=3, 6.0%), Poland (n=2, 4.0%), Canada (n=2, 4.0%), and Italy (n=2, 4.0%).

5.4 Survey Procedure

Upon entering the survey, the participants were first introduced to the study, and consent was requested before proceeding. Next, the participants were given a definition of a persona ("*A persona is an imaginary [fictitious] person that describes a type of user. For example, 'Lazy Larry' could be a persona describing a demotivated worker.*"), and an algorithmically generated persona ("*Algorithmically generated personas are a particular type of persona. They are created automatically from social media data, using data science algorithms.*"). The participants were also introduced to the toxic quote problem and the Hate Filter functionality. After this, the participants were asked to give feedback on the Hate Filter functionality. In addition to the statements given in Table 2 which were answered using a numerical scale, we also asked participants to give open-ended feedback regarding the feature's risks, benefits, and use cases. These open answers were qualitatively analyzed, whereas the numerical answers were analyzed using quantitative statistics.

5.5 Results

The numerical assessment of the Hate Filter is shown in Figure 5. As the Likert scale is widely used as an interval scale in social science research [87], we also treat it as such, reporting means (M) and standard deviations (SD) in our data. The high SDs for most of the perceptions illustrate the different preferences among users when it comes to assessing different aspects of the feature. These differences become salient in the open-ended answers, which we discuss shortly. *Effectiveness* is an exception, as it has a low standard deviation (SD = 0.32), with most participants seeing it as low (M = 2.94). *Usefulness* and *Risk factor* are both high, indicating that the feature is seen to simultaneously have both risks and benefits. Coupled with the high SDs (SD = 1.42 for usefulness and SD = 1.40 for risk factor), this implies that participants are divided, with some siding with the toxicity filtering feature and others perceiving it as a high-risk option. Again, these attitude differences are salient in the open-ended responses. *Trust* is above the mid-range (M = 4.30, SD = 1.33, with the mid-range being 4), indicating that, on average, the participants trust the feature more than they distrust it. However, the participants do not necessarily see the feature as required for portraying user segments, which is shown in the relatively low effectiveness score (M = 2.94, SD = 0.32). *Insights* (M = 3.76, SD = 1.68) and *Realism* (M = 3.62, SD = 1.77) are both below the mid-range, indicating that toxicity filtering is on average seen as taking away some aspects of user insights and the realism of the persona. Also, the participants see the feature as improving their *User experience* of personas, although not dramatically (M = 4.64, SD = 1.32).

Furthermore, we qualitatively categorized the open-ended answers to better understand under which conditions the participants would see the feature as beneficial or harmful. This categorization was carried out using the textual features of standard spreadsheet software. One of the researchers reviewed all of the responses and assigned labels to them that corresponded with their conceptual separation (i.e., different meanings or "themes" [14]). Another researcher reviewed the labels and provided additional insights. Finally, the interpretations were written in the form of a qualitative analysis, as shown below. Overall, this procedure closely corresponds to the thematic analysis approach offered by Clarke et al. [14].

In cases where participants indicated their use of the feature would depend on something, they were asked to specify the issue. The most prominent category in these specifying answers was *topic*, where the participants strongly indicated their use of the feature would depend on what issue they were investigating or writing about (e.g., "It would depend on the topic you are researching for your journalism [sic]" [P05]). If the topic would be controversial, then it might be justified to either turn off toxic quotes so as not to let them become a distraction, or to enable them to see the full picture of audience responses:

- "Depends on whether or not I am writing on a contentious or divisive subject" (P08).

[2] https://www.prolific.co/demographics/

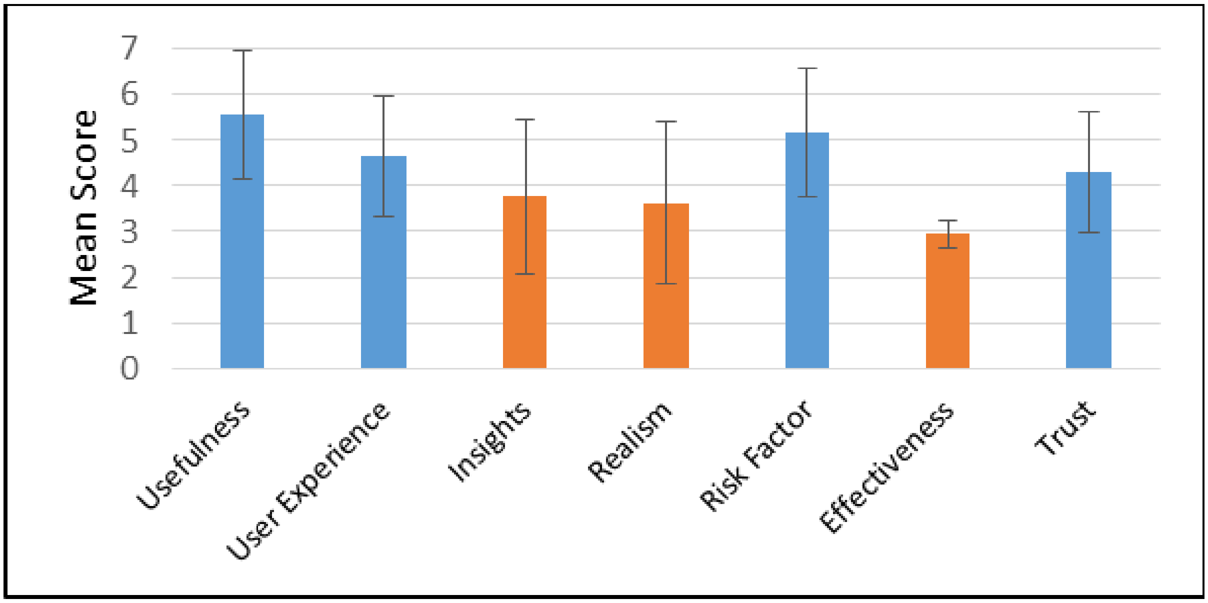


Figure 5: Numerical assessment of the Hate Filter. The bars indicate the mean scores given by the participants; the error bars indicate the standard deviation of the scores.

- "It depends on what my field would be. For example, if I were a music journalist, I would use the filter because hateful comments would not be relevant. If my job was to report about politics, on the other hand, I would want to see all opinions, even if I think they are abhorrent" (P09).
- "Depends on the topic. If it would be sensitive topic for a private person (for example, a mother who lost her children) I'd use that. But if it was news about world/national politics—then no, as it may be interpreted as censorship" (P10).
- "It would depend on whether I wanted general information on the type of people I'd be addressing so I knew how to tailor the content; but if I wanted the true picture for some reason (e.g., if I was looking into toxic online behaviour or bigotry online) I'd probably want to look at it unfiltered" (P13).

The results imply that different topics result in different information needs, where for some cases toxic comments can be useful information, and for others they would not be useful. We identified six specific use cases in the participants' responses to the open-ended question "In what kind of journalistic tasks do you see the toxicity-filtering feature as being useful?" (Table 3). Avoiding an extreme view of the online audiences was the most striking theme, and compatible with the premise proposed in the introduction that "most people are not toxic". Several responses were based on a similar assumption of extreme views distracting from the mainstream (non-toxic) news readers (e.g., "If you're writing fairly light content that is only going to be as controversial by people who tend to be toxic online. Or if you are tailoring content to a specific audience where trolling/bigotry etc [sic] is unlikely to be much of an issue. Then you'd probably be able to look past all the noise and get better insights about what interests people in general" [P30]). Interestingly, P24 referred to specifically learning about toxic audience segments, with the implication of developing a further functionality to only show those personas with toxic quotes. However, the current implementation of PURIFIER works in a different way, filtering toxic quotes *out*, not *in*.

The open-ended responses were particularly insightful concerning the risks of the feature (see Table 4). The biggest concern was *inaccuracy*, i.e., whether the persona no longer matches reality (e.g., "You don't know the true nature of a person" [P4]). Particularly, the participants feared that an omission of toxic quotes might make the personas seem less toxic than they actually are, which would hinder their usefulness for content creation tasks ("If researching topics such as conspiracy theorists, racism, or extreme political positions, the filter could give too moderate an impression" [P8]). The participants used words like "skew" ("Skewing data or viewpoints" [P10]), "false impression" ("Content creators might miss interesting data or gain a false impression of a situation" [P14]), "untruthful" ("It may be 'untruthful' or not painting the full picture regarding someone, or a particular story" [P16]), and "misleading" ("that the values filtered are justified and a misleading persona is created" [P39]) to highlight these concerns.

Incompleteness is similar to inaccuracy but is nonetheless a distinct concept that refers to toxicity filtering hiding certain facets of the persona. Therefore, the persona may be accurate but still be incomplete, as its "dark side" is not revealed to the user. Twelve participants (24.0%) mentioned this risk by way of: "limited understanding or view of topic or persona" (P5); "It can disguise a side of the persona that is very much an important part of it" (P22); "You only see the 'good' of people and omit their true selves" (P45); "You

Table 3: Participant-generated use cases for Hate Filter.

Theme	Example
Fighting hate speech	"Fighting hate speech of any kind." (P03)
Protecting younger audiences	"presenting a persona to a younger audience" (P04)
Stories dealing with sensitive themes	"In articles that deal with highly sensitive and/or controversial themes. For example, recently in my country there was a campaign where people who had experienced abuse as children shared their stories. It was painful to read any toxic/hateful comments after listening to these people. In this case, a toxicity filter would have been useful." (P07)
Research focus	"If the comments are not relevant to the topic then I think the filter can be useful while doing research." (P20)
Avoid getting an extreme view	"For getting a sense of audience, without relying on the 'those who shout loudest get most attention' element. We have a view of our readers who engage - comment on posts etc. But often people post more intensely online than they would in real life and might engage in hyperbole to make a point. Not everyone who makes a comment is as extreme as the comments might imply. This would impact deciding on the kind of stories we cover and how much prominence we give them - are they of interest to our audience?" (P15)
Understanding toxic users	"I think it would be more useful if you actually wanted to find toxic profiles for a specific purpose, e.g. if researching an article on hate speech." (P24)

Table 4: Risks relating to toxicity filtering as observed by the participants.

Category	Frequency of mentions (% of total)
Inaccuracy	16 (32.0%)
Incompleteness	12 (24.0%)
False positives	5 (10.0%)
Filter bubble	4 (8.0%)
Censorship	4 (8.0%)
No risks	4 (8.0%)
Eliminates accountability	2 (4.0%)
Lack of reflection	1 (2.0%)
Aggravation	1 (2.0%)
Language coverage	1 (2.0%)

don't see the in extremis [sic] positions which, when weighed correctly, might show you where moderate users draw the line." (P18). *Filter bubble* refers to blocking out views and information that could be useful or even needed [39], and was mentioned by four (8.0%) participants. One participant mentioned that there might be legitimate concerns in toxic comments that would be left unobserved if such comments were hidden ("If you're entirely unaware of it you might provoke further toxic reactions. Some of what appears to be toxic might actually be people flagging up problems with the content, but just doing it in a dysfunctional/aggressive way" [P30]). *Lack of reflection* is a similar theme, and refers to losing context and insights into how a news story was received and why ("I'd be wary that using the filter could paint an inaccurate picture of the audience. For example, if the majority of the comments were toxic, but the filter covered/removed them, there would be no chance to question exactly why this article/video had attracted this kind of audience." [P7]).

Aggravation refers to leaving audience members unheard and without a voice ("I see some people getting upset with their views not being heard." [P19]). The idea is that ignoring toxic voices would only make them grow louder. Four participants (8.0%) also found toxicity filtering problematic from the perspective of freedom of speech, as toxicity filtering could be interpreted as a form of *censorship* ("censorship or lack of freedom to express ourselves" [P2]). There were two technical concerns that the participants mentioned: *false positives* and *language coverage*. The former refers to the system making an incorrect judgment of a comment's toxicity and therefore operating in an unfair manner ("It may hide legitimate comments." (P37); "AI can be wrong and can't necessarily detect any nuances. It could, for example, mute someone that was quoting someone else, not in agreement, just to provide context." (P20)). The language concern refers to the ability of the system to filter out comments written in different languages ("It may filter out non-toxic comments that just use the same language." [P36]). Indeed, the current implementation of PURIFIER is only compatible with the English language.

Four participants (8.0%) perceived *no risk* to be associated with toxicity filtering (e.g., "I don't see any risk. I am in favor." [P40]; "I don't think there will be any risks" [P6]). These participants could be considered as *Hate-Filter Advocates*, but in contrast, there is a distinct group of participants who can be labeled as *Hate-Filter Critics* (e.g., "It may hide true views and expressions that the personas may have. If you are just open to the positive quotes, you may not fully understand their real persona without seeing their toxic quotes. Seeing the toxic quotes of a persona may change your whole perspective and impression of them and I think that's important." [P32]). This division also shows in regard to the participants' willingness to use the feature. Participants were strongly divided in their views, with almost identical numbers judging the feature positively (n = 17) versus negatively (n = 16). A similar number of participants (n = 17) indicated their use would depend on contextual factors. Moreover, the Pearson's correlation between usefulness and risk factor is $r = -0.39$, $p = .004$, which indicates a statistically significant negative relationship. Notably, participants who give the toxicity filtering feature a high score for usefulness are more likely to give it a low score for riskiness, and *vice versa*.

6 DISCUSSION

6.1 Theoretical and Practical Implications

This research contributes an evaluation of an automated tool for screening (and removing) toxic quotes from algorithmically generated personas. Given the widespread use of personas in HCI research, the topic is relevant to both practitioners and researchers. Particularly, the choices made by algorithms pose a vexing issue for organizations that create social media personas and whose social media comments contain high levels of toxicity, such as online news channels that often garner toxic comments [46, 59]. This issue is connected to the broader picture of moderation challenges in social media [24, 31], namely, *to what extent should moderation of comment datasets be done and by whom?*

Using data science algorithms for persona creation implies that automatically creating social media personas drastically differs from the traditional persona creation, where human persona creators have complete control over what quotes to include in the persona profiles, usually hand-picking comments that are not toxic or writing the comments themselves. Thus, the use of algorithms shifts some of this persona creation *agency* from humans to the algorithm that has no consideration for culturally or socially sensitive topics, referred to as the lack of design sensibilities [26]. We addressed this issue by automatically labeling the persona quotes for toxicity in a background system, and then giving the persona users a choice of either showing or hiding the toxic quotes. Doing so is more suitable than automatically hiding or deleting of the comments, which would violate the accuracy principle of personas (i.e., misleading decision-makers in their audience), as well as compromising the freedom of expression (i.e., blocking out dissident views) with the negative effect of harming the content creators' perceptions of personas.

The aspect of *choice* is also important here because it seems to align with our findings. Namely, there is a group of content creators that favors the toxicity filtering functionality (roughly one-third of the participants). There is another group that is skeptical about the functionality (again, roughly one-third of the participants). Finally, there is a group that sees the value of functionality as being tied to the context of a content creation task (the last third of the participants). In particular, the topic the content creator is researching or writing about impacts the perceived usefulness of toxicity filtering. For example, some investigative content creators may purposefully seek to understand hate communities or to discover misinformation online [44], and for them, having an unfiltered view of the comments is a necessary requirement for their work. Conceptually, this is associated with the goal of creating extreme personas (or characters [19]), and specifically focusing on deviant outlier behaviors. It was also interesting to observe that the participants' responses mirrored the concerns pointed out in previous literature. Namely, the fact that altering the quotes may increase the perception of personas being misleading matters, as 'misleading' is specifically mentioned [43] as one of the negative stereotypes that users associate with personas (albeit there being a risk that personas are always stereotypical to some extent [83]). Hence, this connotation of biasing the user representation is not desirable for all stakeholders. Therein lies an interesting distinction and conflict of biases; namely, that toxic quotes can bias the stakeholder perception towards negativity, but removing the toxicity can bias it to positivity. Both of these types of bias can have unintended consequences for design. One option to explore here is to more strongly convey to stakeholders that personas are a *group* of people [27], and though this, possibly limit the biasing effect of mixed sentiments in the persona quotes.

From an ethical point of view, if the PURIFIER would be enabled by default, then two fundamental rights could be violated – those of *people that the personas are based on*, and those of *people using the personas*. The former would be susceptible to censorship, muting, and "canceling," which cannot be seen as optimal solutions to the toxicity problem [12], and the latter would be misguided by giving them a too "rosy" picture of their real audience. The automatic exclusion of toxic comments from personas could even be considered as a dark pattern of design, i.e., a UI design choice that steers or deceives users into making potentially harmful decisions [42]. Hence, from a system development perspective, we postulate that the solution is to give content creators the *option* to hide or not hide toxic comments, rather than the system making this choice on their behalf. The value of this functionality is that content creators can use it to show only "serious" (constructive) comments, instead of being upset or derailed by toxic comments. As shown by the strong division in the participants' attitudes for toxicity filtering, there are those who keenly advocate the feature and those who show a great deal of hesitancy. Therefore, there is a need for user choice in "co-creating" the personas with algorithms, which the Hate Filter exemplifies.

6.2 Limitations and Future Research and Development

This study has some limitations that the reader should be aware of. First, the participants were screened to be working in publishing. While this was the closest option afforded by the data collection platform, publishing could be seen as being much broader than online content creation. We did not specifically gauge the participants' experience with online content creation, which is something that should be done in future work. On the other hand, we also did not observe any confusion or misunderstanding in the open answers given by the participants that would indicate them as having not understood the point of using toxicity filtering for content creation with personas.

Second, the questionnaire contained concepts that might not be known by participants, such as UX. Again, while we did not observe any difficulties based on the participants' open answers, the questions may leave some room for interpretation. Furthermore, the underlying constructs that the questions aimed to evaluate might not have been fully captured, even though, again, from the open answers, confusion among the participants seems unlikely.

Overall, social media persona creation is challenging because a persona, by definition as representing a 'group of people', contains diverse views. Hence, in a strict sense, most social media personas would inevitably have both less and more toxic facets. One of the participants expressed this same idea: "Sadly, sometimes there IS an extreme element - ideally you want something that will reflect the general feeling but not overlook the vocal minority." [P15]). Given this heterogeneity, future work could investigate a mixture of toxic and non-toxic quotes in AGPs, answering questions such as: Are there thresholds to frequency and strength of toxicity that can "tip over" the users' perceptions of the persona? Is the "one rotten apple spoils the basket" effect valid, or would persona users be able to understand that a single toxic quote does not represent all users corresponding with the persona? These issues need more investigation.

Future work could also address the saliency and active use of toxicity filtering in interactive persona systems. Our data shows that users are not actively applying the feature in a naturalistic setting, which could be for a number of reasons. For example, the users might not perceive the feature in the UI (saliency issue) or properly understand how it works (on-boarding or explainability issue), or they simply might not want to hide toxic quotes. These aspects inhibiting the use of toxicity filtering should be clarified in future research. Finally, two specific development ideas were inferred from the user feedback: (1) *creation of purely toxic personas* (i.e., "filtering in" toxic quotes) for special focus on toxic personas, (2) and *increasing PURIFIER's language coverage from English to other languages* which is important because social media content often attracts comments in various languages. PURIFIER should incorporate these features in the future, and there is also a need for removing bot-generated comments from the source material, as bots play a "social role" in algorithms that are trained using social media data [76], including the current use case of creating social media personas.

Finally, there are concerns of ethical persona creation [60], which could be defined as personas that do not violate the rights of any stakeholder groups. While we advise persona creators to apply toxicity *detection* when creating personas from social media user data, this does not automatically mean the *censoring* of toxic quotes from AGPs by default (as the latter would be using algorithmic power to make choices on information seen by users on their behalf). The matter of toxicity is more complex than that, and involves at least two major ethical considerations. First, there is an issue of truthfulness or accuracy. If, indeed, many comments in the channel's data are toxic, then content creators that are using personas to understand their audiences need to be made aware of this. In other words, is it not the *right* of the content creator to know about the "dark side" of their audience? Second, toxicity filtering may affect the voice of social media users[17], prompting the question as to whether toxic user types have the *right* to be represented in social media personas? In a broader sense, this discussion relates to a larger theme in using algorithms to battle toxicity. That is, *how should toxicity classification scores be used in real systems?* Previous research in HCI has featured surprisingly little discussion on this matter, even though toxicity is regarded as a serious matter of human interaction via computing systems [22, 78, 82]. Therefore, a discussion on the normative basis of dealing with toxicity in computing systems is needed. Specifically: *What is the standpoint of HCI on deploying toxicity detection in computing systems?*

7 CONCLUSION

In the absence of human supervision, algorithmically generated personas may be overwhelmed with toxic quotes. As these toxic quotes may have detrimental effects on content creators' perceptions of the personas, there is a need to provide tools for content creators to manage such comments. We demonstrated the integration of such functionality to an algorithmic persona generation system, while discussing its impact on ethical persona creation, considering both pros and cons. A feasibility study revealed both hate-filter critics (~1/3 of the participants), advocates (~1/3), as well as those who fell in between (~1/3). The usefulness of toxicity filtering is primarily driven in the online content creation context, by the type of stories content creators are writing, and for some topics, toxic quotes can be useful information, while for others, they distract or mislead. Therefore, the control of toxicity filtering should be left in the hands of the end users of the personas and users of the persona systems.

REFERENCES

[1] Kamil Akhuseyinoglu and Peter Brusilovsky. 2021. Data-Driven Modeling of Learners' Individual Differences for Predicting Engagement and Success in Online Learning. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization.* Association for Computing Machinery, New York, NY, USA, 201–212. Retrieved January 19, 2022 from https://doi.org/10.1145/3450613.3456834

[2] Nalini Ambady and Robert Rosenthal. 1992. Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychol. Bull.* 111, 2 (1992), 256–274. DOI:https://doi.org/10.1037/0033-2909.111.2.256

[3] Christos Amyrotos. 2021. Adaptive Visualizations for Enhanced Data Understanding and Interpretation. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization.* Association for Computing Machinery, New York, NY, USA, 291–297. Retrieved January 19, 2022 from https://doi.org/10.1145/3450613.3459657

[4] Jisun An, Haewoon Kwak, Soon-gyo Jung, Joni Salminen, and Bernard J. Jansen. 2018. Customer segmentation using online platforms: isolating behavioral and demographic segments for persona creation via aggregated user data. *Soc. Netw. Anal. Min.* 8, 1 (2018). DOI:https://doi.org/10.1007/s13278-018-0531-0

[5] Jisun An, Haewoon Kwak, Joni Salminen, Soon-gyo Jung, and Bernard J. Jansen. 2018. Imaginary People Representing Real Numbers: Generating Personas from Online Social Media Data. *ACM Trans. Web TWEB* 12, 3 (2018).

[6] Farshid Anvari, Deborah Richards, Michael Hitchens, and Muhammad Ali Babar. 2015. Effectiveness of Persona with Personality Traits on Conceptual Design. In *Proceedings of the 37th International Conference on Software Engineering - Volume 2* (ICSE '15), IEEE Press, Piscataway, NJ, USA, 263–272. Retrieved July 11, 2018 from http://dl.acm.org/citation.cfm?id=2819009.2819048

[7] Farshid Anvari, Deborah Richards, Michael Hitchens, Muhammad Ali Babar, Hien Minh Thi Tran, and Peter Busch. 2017. An empirical investigation of the influence of persona with personality traits on conceptual design. *J. Syst. Softw.* 134, (December 2017), 324–339. DOI:https://doi.org/10.1016/j.jss.2017.09.020

[8] Farshid Anvari, Deborah Richards, Michael Hitchens, and Hien Minh Thi Tran. 2019. Teaching User Centered Conceptual Design Using Cross-Cultural Personas and Peer Reviews for a Large Cohort of Students. In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET)*, 62–73. DOI:https://doi.org/10.1109/ICSE-SEET.2019.00015

[9] Ahmer Arif, Leo Graiden Stewart, and Kate Starbird. 2018. Acting the part: Examining information operations within# BlackLivesMatter discourse. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (2018), 1–27.

[10] Asa Blomquist and Mattias Arvola. 2002. Personas in action: ethnography in an interaction design team. In *Proceedings of the second Nordic conference on Human-computer interaction*, ACM, Aarhus, Denmark, 197–200. Retrieved May 28, 2017 from http://dl.acm.org/citation.cfm?id=572044

[11] Susanne Bødker, Ellen Christiansen, Tom Nyvang, and Pär-Ola Zander. 2012. Personas, people and participation: challenges from the trenches of local government. In *Proceedings of the 12th Participatory Design Conference: Research Papers-Volume 1*, 91–100.

[12] Gwen Bouvier. 2020. Racist call-outs and cancel culture on Twitter: The limitations of the platform's ability to define issues of social justice. *Discourse Context Media* 38, (2020), 100431.

[13] Chris Chapman and Russell P. Milham. 2006. The Personas' New Clothes: Methodological and Practical Arguments against a Popular Method. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 50, 5 (October 2006), 634–636. DOI:https://doi.org/10.1177/154193120605000503

[14] Victoria Clarke, Virginia Braun, and Nikki Hayfield. 2015. Thematic analysis. *Qual. Psychol. Pract. Guide Res. Methods* 222, 2015 (2015), 248.

[15] Francisco Conejo. 2017. Improving social media brand personas using archetypes. *J. Digit. Soc. Media Mark.* 5, 2 (2017), 189–202.

[16] Alan Cooper. 1999. *The Inmates Are Running the Asylum: Why High Tech Products Drive Us Crazy and How to Restore the Sanity* (1 edition ed.). Sams - Pearson Education, Indianapolis, IN.

[17] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of Eleventh International AAAI Conference on Web and Social Media*, Montreal, Canada, 512–515.

[18] [18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv Prepr. ArXiv181004805* (2018).

[19] [19] John Partomo Djajadiningrat, William W. Gaver, and J. W. Fres. 2000. Interaction relabelling and extreme characters: methods for exploring aesthetic interactions. In *Proceedings of the 3rd conference on Designing interactive systems: processes, practices, methods, and techniques*, 66–71.

[20] [20] Motahhare Eslami, Sneha R. Krishna Kumaran, Christian Sandvig, and Karrie Karahalios. 2018. Communicating Algorithmic Process in Online Behavioral Advertising. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ACM, Montréal, Canada, 432.

[21] Bruna Moraes Ferreira, Simone DJ Barbosa, and Tayana Conte. 2016. Pathy: Using empathy with personas to design applications that meet the users' needs. In *International Conference on Human-Computer Interaction*, Springer, 153–165.

[22] Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv. CSUR* 51, 4 (2018), 1–30.

[23] Simson Garfinkel, Jeanna Matthews, Stuart S. Shapiro, and Jonathan M. Smith. 2017. Toward Algorithmic Transparency and Accountability. *Commun ACM* 60, 9 (August 2017), 5–5. DOI:https://doi.org/10.1145/3125780

[24] Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data Soc.* 7, 1 (January 2020), 2053951719897945. DOI:https://doi.org/10.1177/2053951719897945

[25] Jonathan Grudin. 2006. Why Personas Work: The Psychological Evidence. In *The Persona Lifecycle*, John Pruitt and Tamara Adlin (eds.). Elsevier, 642–663. DOI:https://doi.org/10.1016/B978-012566251-2/50013-7

[26] Jonna Häkkilä, Mikael Wiberg, Nils Johan Eira, Tapio Seppänen, Ilkka Juuso, Maija Mäkikalli, and Katrin Wolf. 2020. Design Sensibilities-Designing for Cultural Sensitivity. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*, 1–3.

[27] Charles G. Hill, Maren Haag, Alannah Oleson, Chris Mendez, Nicola Marsden, Anita Sarma, and Margaret Burnett. 2017. Gender-Inclusiveness Personas vs. Stereotyping: Can We Have it Both Ways? In *Proceedings of the 2017 CHI Conference*, ACM Press, 6658–6671. DOI:https://doi.org/10.1145/3025453.3025609

[28] Richard J. Holden, Carly N. Daley, Robin S. Mickelson, Davide Bolchini, Tammy Toscos, Victor P. Cornet, Amy Miller, and Michael J. Mirro. 2020. Patient decision-making personas: An application of a patient-centered cognitive task analysis (P-CTA). *Appl. Ergon.* 87, (September 2020), 103107. DOI:https://doi.org/10.1016/j.apergo.2020.103107

[29] Aaron Humphrey. 2017. User Personas and Social Media Profiles. *Pers. Stud.* 3, 2 (December 2017), 13–20.

[30] Bernard Jansen, Joni Salminen, Soon-gyo Jung, and Kathleen Guan. 2021. *Data-Driven Personas* (1st ed.). Morgan & Claypool Publishers.

[31] Jialun Aaron Jiang, Charles Kiene, Skyler Middler, Jed R. Brubaker, and Casey Fiesler. 2019. Moderation challenges in voice-based online communities on discord. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (2019), 1–23.

[32] Soon-Gyo Jung, Joni Salminen, Jisun An, Haewoon Kwak, and B. J. Jansen. 2018. Automatically Conceptualizing Social Media Analytics Data via Personas. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM 2018)*, San Francisco, California, USA.

[33] Soon-Gyo Jung, Joni Salminen, and Bernard J. Jansen. 2019. Personas Changing Over Time: Analyzing Variations of Data-Driven Personas During a Two-Year Period. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI EA '19), ACM, New York, NY, USA, LBW2714:1-LBW2714:6. DOI:https://doi.org/10.1145/3290607.3312955

[34] Soon-gyo Jung, Joni Salminen, and Bernard J. Jansen. 2021. All About the Name: Assigning Demographically Appropriate Names to Data-Driven Entities. In *Proceedings of the Hawaii International Conference on System Sciences (HICSS2021)*, Virtual conference.

[35] Soon-Gyo Jung, Joni Salminen, Haewoon Kwak, Jisun An, and Bernard J. Jansen. 2018. Automatic Persona Generation (APG): A Rationale and Demonstration. In *Proceedings of the 2018 Conference on Human Information Interaction&Retrieval*, ACM, 321–324.

[36] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T. Hancock, and Michael S. Bernstein. 2020. Conceptual metaphors impact perceptions of human-AI collaboration. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2 (2020), 1–26.

[37] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 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* (CHI '09), ACM, New York, NY, USA, 1509–1512. DOI:https://doi.org/10.1145/1518701.1518930

[38] Daniel D. Lee and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature* 401, 6755 (October 1999), 788–791. DOI:https://doi.org/10.1038/44565

[39] Q. Vera Liao and Wai-Tat Fu. 2013. Beyond the filter bubble: interactive effects of perceived threat and topic involvement on selective exposure to information. In *Proceedings of the SIGCHI conference on human factors in computing systems*, 2359–2368.

[40] Caitlin Lustig. 2019. Intersecting imaginaries: visions of decentralized autonomous systems. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (2019), 1–27.

[41] Nicola Marsden and Maren Haag. 2016. Stereotypes and Politics: Reflections on Personas. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (CHI '16), ACM, New York, NY, USA, 4017–4031. DOI:https://doi.org/10.1145/2858036.2858151

[42] Arunesh Mathur, Gunes Acar, Michael J. Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark patterns at scale: Findings from a crawl of 11K shopping websites. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (2019), 1–32.

[43] Tara Matthews, Tejinder Judge, and Steve 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* (CHI '12), ACM, New York, NY, USA, 1219–1228. DOI:https://doi.org/10.1145/2207676.2208573

[44] Melinda McClure Haughey, Meena Devii Muralikumar, Cameron A. Wood, and Kate Starbird. 2020. On the Misinformation Beat: Understanding the Work of Investigative Journalists Reporting on Problematic Information Online. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2 (2020), 1–22.

[45] Jennifer Jen McGinn and Nalini Kotamraju. 2008. Data-driven persona development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, 1521–1524.

[46] Yelena Mejova, Amy X. Zhang, Nicholas Diakopoulos, and Carlos Castillo. 2014. Controversy and Sentiment in Online News. In *arXiv:1409.8152 [cs]*, Miami, FL, USA. Retrieved September 17, 2019 from http://arxiv.org/abs/1409.8152

[47] T. Mijač, M. Jadrić, and M. Ćukušić. 2018. The potential and issues in data-driven development of web personas. In *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 1237–1242. DOI:https://doi.org/10.23919/MIPRO.2018.8400224

[48] L. Molenaar. 2017. Data-driven personas: Generating consumer insights with the use of clustering analysis from big data. Master Thesis. TU Delft, Netherlands. Retrieved March 17, 2018 from http://resolver.tudelft.nl/uuid:12d7f261-20b4-4656-93d7-fed2b437aefb

[49] Lene Nielsen. 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* (DIS '02), ACM, New York, NY, USA, 99–104. DOI:https://doi.org/10.1145/778712.778729

[50] Lene Nielsen. 2019. *Personas - User Focused Design* (2nd ed. 2019 edition ed.). Springer, New York, NY.

[51] Lene Nielsen, Kira Storgaard Hansen, Jan Stage, and Jane Billestrup. 2015. A Template for Design Personas: Analysis of 47 Persona Descriptions from Danish Industries and Organizations. *Int. J. Sociotechnology Knowl. Dev.* 7, 1 (2015), 45–61. DOI:https://doi.org/10.4018/ijskd.2015010104

[52] Lene Nielsen, Soon-Gyo Jung, Jisun An, Joni Salminen, Haewoon Kwak, and Bernard 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* (OZCHI '17), ACM, New York, NY, USA, 602–606. DOI:https://doi.org/10.1145/3152771.3156178

[53] Lene Nielsen and Kira 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, 1665–1674.

[54] Farzana Parveen, Noor Ismawati Jaafar, and Sulaiman Ainin. 2015. Social media usage and organizational performance: Reflections of Malaysian social media managers. *Telemat. Inform.* 32, 1 (2015), 67–78.

[55] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

[56] Alisha Pradhan, Leah Findlater, and Amanda Lazar. 2019. " Phantom Friend" or" Just a Box with Information" Personification and Ontological Categorization of Smart Speaker-based Voice Assistants by Older Adults. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (2019), 1–21.

[57] John Pruitt and Jonathan Grudin. 2003. Personas: Practice and Theory. In *Proceedings of the 2003 Conference on Designing for User Experiences* (DUX '03), ACM, New York, NY, USA, 1–15. DOI:https://doi.org/10.1145/997078.997089

[58] Psychology Research and Reference. 2018. Person Perception. Retrieved August 28, 2018 from https://psychology.iresearchnet.com/social-psychology/social-cognition/person-perception/

[59] Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard J. Jansen. 2018. 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 (ICWSM 2018)*, San Francisco, California, USA. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17885

[60] Joni Salminen, Willemien Froneman, Soon-gyo Jung, Shammur Chowdhury, and Bernard J. Jansen. 2020. The Ethics of Data-Driven Personas. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems Extended Abstracts* (CHI '20), Association for Computing Machinery, Honolulu, HI, USA, 1–9. DOI:https://doi.org/10.1145/3334480.3382790

[61] Joni Salminen, Kathleen Guan, Soon-Gyo Jung, and Bernard J. Jansen. 2021. A Survey of 15 Years of Data-Driven Persona Development. *Int. J. Human–Computer Interact.* 0, 0 (April 2021), 1–24. DOI:https://doi.org/10.1080/10447318.2021.1908670

[62] Joni Salminen, Maximilian Hopf, Shammur A. Chowdhury, Soon-gyo Jung, Hind Almerekhi, and Bernard J. Jansen. 2020. Developing an online hate classifier for multiple social media platforms. *Hum.-Centric Comput. Inf. Sci.* 10, 1 (2020), 1. DOI:https://doi.org/10.1186/s13673-019-0205-6

[63] Joni Salminen, Bernard J. Jansen, Jisun An, Haewoon Kwak, and Soon-gyo Jung. 2018. Are personas done? Evaluating their usefulness in the age of digital analytics. *Pers. Stud.* 4, 2 (November 2018), 47–65. DOI:https://doi.org/10.21153/psj2018vol4no2art737

[64] Joni Salminen, Bernard J. Jansen, Jisun An, Haewoon Kwak, and Soon-Gyo Jung. 2019. Automatic Persona Generation for Online Content Creators: Conceptual Rationale and a Research Agenda. In *Personas - User Focused Design*, Lene Nielsen (ed.). Springer London, London, 135–160. DOI:https://doi.org/10.1007/978-1-4471-7427-1_8

[65] [65] Joni Salminen, Soon-Gyo Jung, Jisun An, Haewoon Kwak, and Bernard J. Jansen. 2018. Findings of a User Study of Automatically Generated Personas. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems* (CHI EA '18), ACM, New York, NY, USA, LBW097:1-LBW097:6. DOI:https://doi.org/10.1145/3170427.3188470

[66] Joni Salminen, Soon-gyo Jung, Jisun An, Haewoon Kwak, Lene Nielsen, and Bernard J. Jansen. 2019. Confusion and information triggered by photos in persona profiles. *Int. J. Hum.-Comput. Stud.* 129, (September 2019), 1–14. DOI:https://doi.org/10.1016/j.ijhcs.2019.03.005

[67] Joni Salminen, Soon-gyo Jung, Shammur Absar Chowdhury, Sercan Sengün, and Bernard J Jansen. 2020. Personas and Analytics: A Comparative User Study of Efficiency and Effectiveness for a User Identification Task. In *Proceedings of the ACM Conference of Human Factors in Computing Systems (CHI'20)*, ACM, Honolulu, Hawaii, USA. DOI:https://doi.org/10.1145/3313831.3376770

[68] Joni Salminen, Soon-Gyo Jung, João Santos, and Bernard J. Jansen. 2021. Toxic Text in Personas: An Experiment on User Perceptions. *AIS Trans. Hum.-Comput. Interact.* 13, 4 (December 2021), 453–478. DOI:https://doi.org/10.17705/1thci.00157

[69] Joni Salminen, Soon-gyo Jung, João M. Santos, and Bernard J. Jansen. 2019. Does a Smile Matter if the Person Is Not Real?: The Effect of a Smile and Stock Photos on Persona Perceptions. *Int. J. Human–Computer Interact.* 0, 0 (September 2019), 1–23. DOI:https: //doi.org/10.1080/10447318.2019.1664068

[70] Joni Salminen, Soon-gyo Jung, João M. Santos, Ahmed Mohamed Kamel, and Bernard J. Jansen. 2021. Picturing It!: The Effect of Image Styles on User Perceptions of Personas. In *In the Proceedings of ACM Human Factors in Computing Systems (CHI'21)*, ACM, Virtual conference.

[71] Joni Salminen, Lene Nielsen, Soon-Gyo Jung, Jisun An, Haewoon Kwak, and Bernard J Jansen. 2018. "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)*, Montréal, Canada.

[72] Joni Salminen, Joao M. Santos, Soon-gyo Jung, Motahhare Eslami, and Bernard J. Jansen. 2019. Persona Transparency: Analyzing the Impact of Explanations on Perceptions of Data-Driven Personas. *Int. J. Human–Computer Interact.* 0, 0 (November 2019), 1–13. DOI:https://doi.org/10.1080/10447318.2019.1688946

[73] Joni Salminen, Joao M. Santos, Haewoon Kwak, Jisun An, Soon-gyo Jung, and Bernard J. Jansen. 2020. Persona Perception Scale: Development and Exploratory Validation of an Instrument for Evaluating Individuals' Perceptions of Personas. *Int. J. Hum.-Comput. Stud.* 141, (April 2020), 102437. DOI:https://doi.org/10.1016/j.ijhcs.2020.102437

[74] Joni Salminen, Sercan Sengün, Juan Corporan, Soon-gyo Jung, and Bernard J. Jansen. 2020. Topic-driven toxicity: Exploring the relationship between online toxicity and news topics. *PLOS ONE* 15, 2 (February 2020), e0228723. DOI:https://doi.org/10.1371/journal.pone.0228723

[75] Joni Salminen, Sercan Şengün, Haewoon Kwak, Bernard J. Jansen, Jisun An, Soon-gyo Jung, Sarah Vieweg, and Fox Harrell. 2018. From 2,772 segments to five personas: Summarizing a diverse online audience by generating culturally adapted personas. *First Monday* 23, 6 (June 2018). Retrieved June 3, 2018 from http://firstmonday.org/ojs/index.php/fm/article/view/8415

[76] Joseph Seering, Juan Pablo Flores, Saiph Savage, and Jessica Hammer. 2018. The social roles of bots: evaluating impact of bots on discussions in online communities. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (2018), 1–29.

[77] Cathrine Seidelin, A. Jonsson, M. Høgild, J. Rømer, and P. Diekmann. 2014. Implementing personas for international markets: a question of UX maturity. In *Proceedings at SIDER.*

[78] Sercan Sengün, Joni Salminen, Peter Mawhorter, Soon-gyo Jung, and B. J. Jansen. 2019. Exploring the Relationship Between Game Content and Culture-based Toxicity: A Case Study of League of Legends and MENA Players. In *In the Proceedings of the 30th ACM Conference on Hypertext and Social Media (HT'19)*, Hof, Germany.

[79] Dimitris Spiliotopoulos, Dionisis Margaris, and Costas Vassilakis. 2020. Data-Assisted Persona Construction Using Social Media Data. *Big Data Cogn. Comput.* 4, 3 (September 2020), 21. DOI:https://doi.org/10.3390/bdcc4030021

[80] Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (2019), 1–26.

[81] Phillip Douglas Stevenson and Christopher Andrew Mattson. 2019. The Personification of Big Data. *Proc. Des. Soc. Int. Conf. Eng. Des.* 1, 1 (July 2019), 4019–4028. DOI:https: //doi.org/10.1017/dsi.2019.409

[82] Selen Türkay, Jessica Formosa, Sonam Adinolf, Robert Cuthbert, and Roger Altizer. 2020. See No Evil, Hear No Evil, Speak No Evil: How Collegiate Players Define, Experience and Cope with Toxicity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.

[83] Phil Turner and Susan Turner. 2011. Is stereotyping inevitable when designing with personas? *Des. Stud.* 32, 1 (2011), 30–44.

[84] Sophie Welber, Valerie Zhao, Claire Dolin, Olivia Morkved, Henry Hoffmann, and Blase Ur. 2021. Do Users Have Contextual Preferencesfor Smartphone Power Management? In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization.* Association for Computing Machinery, New York, NY, USA, 44–54. Retrieved January 19, 2022 from https: //doi.org/10.1145/3450613.3456813

[85] Mark E. Whiting, Grant Hugh, and Michael S. Bernstein. 2019. Fair work: Crowd work minimum wage with one line of code. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 197–206.

[86] Eva Yiwei Wu, Emily Pedersen, and Niloufar Salehi. 2019. Agent, Gatekeeper, Drug Dealer: How Content Creators Craft Algorithmic Personas. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (November 2019), 1–27. DOI:https://doi.org/10.1145/3359321

[87] Huiping Wu and Shing-On Leung. 2017. Can Likert Scales be Treated as Interval Scales?—A Simulation Study. *J. Soc. Serv. Res.* 43, 4 (August 2017), 527–532. DOI:https: //doi.org/10.1080/01488376.2017.1329775

[88] Xiang Zhang, Hans-Frederick Brown, and Anil 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* (CHI '16), ACM, New York, NY, USA, 5350–5359.