Evaluating the Social Media Profiles of Online Harassers: An Experimental Study of Attention and Attitudes

SARITA SCHOENEBECK, University of Michigan, USA
YU YIN SHEN, University of Michigan, USA
JILL DAVIDSON, University of Michigan, USA

Online harassment is pervasive. While substantial research has examined the nature of online harassment and how to moderate it, little work has explored how social media users evaluate the profiles of online harassers. This is important for helping people who may be experiencing or observing harassment to quickly and efficiently evaluate the user doing the harassing. We conducted a lab experiment (N=45) that eye-tracked participants while they viewed profiles of users who engaged in online harassment on mock Facebook, Twitter, and Instagram profiles. We evaluated what profile elements they looked at and for how long relative to a control group, and their qualitative attitudes about harasser profiles. Results showed that participants look at harassing users' post history more quickly than non-harassing users. They are also somewhat more likely to recall harassing profiles than non-harassing profiles. However, they do not spend more time on harassing profiles. Understanding what users pay attention to and recall may offer new design opportunities for supporting people who experience or observe harassment online.

CCS Concepts: • **Human-centered computing → Empirical studies in collaborative and social computing**.

Additional Key Words and Phrases: online harassment; attention; social media profiles; eye-tracking; interfaces

ACM Reference Format:
Sarita Schoenebeck, Yu Yin Shen, and Jill Davidson. 2023. Evaluating the Social Media Profiles of Online Harassers: An Experimental Study of Attention and Attitudes. *Proc. ACM Hum.-Comput. Interact.* 7, GROUP, Article 7 (January 2023), 14 pages. https://doi.org/10.1145/3567557

1 INTRODUCTION

Harassment is widespread online and can include insults, shaming, embarrassment, hate speech, doxxing, non-consensual image sharing, and many other harmful behaviors [29, 51, 65]. These behaviors can be one-time events or they can take the form of sustained harassment over long periods [39]. Online harassment has been documented across platforms, including social media like Twitter and Instagram, game platforms like Xbox and Twitch, in journalism, and elsewhere [13, 21, 22, 42, 43]. While the harms to harassment targets are not always visible or known, they can include psychological harm, economic or financial harm, physical harm, relational harm, self-censorship, and chilling effects [2, 20, 24, 33, 39, 47, 64, 65]. Victims of harassment report being worried, scared, angry and/or annoyed by their experience, and these experiences can make them hesitant to engage or post further [51]. While extensive research has examined the nature of online

Authors' addresses: Sarita Schoenebeck, University of Michigan, USA; Yu Yin Shen, University of Michigan, USA; Jill Davidson, University of Michigan, USA.

2573-0142/2023/1-ART7 $15.00
https://doi.org/10.1145/3567557

harassment (e.g. [13, 14, 22, 42, 46, 53]) and how to moderate content (e.g. [28, 36, 49, 63, 65]), less work has focused on the design of social media profiles in the context of online harassment.

Profiles are central to the social media user experience. People scroll through hundreds or thousands of posts from other users daily in their social media newsfeeds. Forming assessments of other people is a fundamental social behavior and people form impressions of others mere seconds after meeting them [5]. This behavior may be magnified and exacerbated on social media, especially on apps like TikTok where people may scroll through 10s or 100s of posts in just a few minutes. In online contexts, social media users rely on elements like photos, friend networks, and content when evaluating others (e.g., [8, 25, 30]); however, there is little known about how the evaluate the profiles of people who engage in harassment. Exposure to harassing content and users then involves ongoing evaluation of that content and those users, making quick assessments about whether to reply, react, ignore, keep scrolling, or block. Theories of attention generally converge on the belief that attention is limited by human processing capacity but there is wide divergence on how filtering of information occurs [58, 60, 71].

These behaviors have been difficult to study because passive behaviors like viewing [18] are not typically available in log file data. They are also difficult to capture in user studies, where people may not also be able to reflect back on what they paid attention to on profiles at any moment in time [11, 67]. Approximations of attention are typically measured via social media system data (e.g., likes, retweets, upvotes, posts, or comments) or via self-reports of behavior, but these may give less insight into passive behaviors like scrolling, viewing content, or evaluating others [31].

We conducted a within-subjects lab experiment (N=45) to examine how participants evaluate user profiles. Participants viewed a harassing or non-harassing (control) comment on mockups we created of Facebook, Instagram, and Twitter profiles then viewed the profile of the commenter. We eye-tracked participants and conducted a cognitive walkthrough and interview. We find that participants shift attention to harassing users' post history more quickly than to non- harassing users. They are also more likely to recall harassing profiles than non- harassing profiles. However, they do not spend more time on harassing profiles.

This study makes a few contributions over prior work. First, it reveals post histories as a potentially important design feature for supporting people who experience or observe harassment. Understanding what profile information people pay attention to is important for designing interfaces that better support social media users' needs and for understanding how people experience and react to harassers. Second, it suggests cognitive costs associated with online harassment via recall.

1.1 Online Harassment

Social media platforms typically rely on community guidelines that indicate what kinds of content is appropriate on their platforms. Platforms govern content through a process called content moderation, where content passes through automated filters and, if flagged or manually reported, through manual filters [36, 63, 68]. The governance process requires multiple phases, including deciding whether content violates guidelines and then if so, deciding what sanction should be enacted. Automated approaches to content moderation use machine learning models that try to identify content that violates guidelines [22]. These models generally perform well on some kinds of spam and on extremely harmful content like child sexual abuse content; but they tend to struggle with more nuanced content like humor and jokes or in-group communication that might seem inappropriate to a machine but is appropriate in the context it is being used [45, 52, 72]. As a result, some content flagged by a machine learning model may result in false positives (where appropriate content is deemed as in violation of guidelines) or false negatives (where violating content is overlooked). Manual content moderation is performed by human workers who review individual pieces of content and similarly try to assess whether the content violated guidelines.

These human workers tend to be underpaid, asked to review psychologically traumatizing content (e.g. violent images), and are outsourced workers via third party firms without the benefits of employees [63].

People with vulnerable and marginalized identities experience substantial harassment online, such as gender minorities, Black, Brown, and Indigenous people, people with disabilities, religious minorities, dissidents, and LGBTQ people [29, 42–44, 51, 65]. Content moderation processes may further disproportionately harm those groups if guidelines are not applied equitably for users [44]. When content is found to violate community guidelines, platforms typically rely on a few types of sanctions [40]. One is a warning to the account owner, which is used for relatively mild violations and first time offenders. Another is removal of the offending content with a notification to the account owner, either temporarily or permanently. Other possible sanctions include "shadow-banning"—an opaque algorithmic response that decreases the likelihood of content showing up in other users' news feeds [7, 37]. These remedies are generally punitive in nature—they punish accounts for violating guidelines rather than seeking to rehabilitate them [66]. While punishment may be an effective response, it is also possible that restorative responses like apologies, mediation, or compensation could be effective alternatives [66]. Shame-based approaches have also been proposed as a form of gender justice [70]. A second challenge with content removal remedies is they introduce concerns about rights to freedom of expression and censorship [24, 65]. Currently, most online harassment research has focused on how to identify and discourage harassment or report it more effectively (e.g. [22, 46, 55]). Sites currently require individual users to manage harassment in their feeds, either by muting, blocking, reporting, or ignoring content. However, this can be burdensome for users who have to repeatedly engage with harassing content and users; designing victim-centered interfaces might reduce that burden.

1.2 Attention and Design

Navigating the social world requires that people adapt, flexibly, to the people they see around them. Attention refers to the process of concentrating on some aspects of information and discarding others [58]. Theories of attention seek to explain what people attend to, why, and how. What a person is looking at, and what they look at first, is thought to be an indicator of "on top of the stack" cognitive processes [48]. In contrast, fixations—where the eyes are stationary—are periods where the brain is encoding or processing information. Eye fixations, where people look at an element, signal that information is extracted about that element and may be entering a cognitive processing phase [48, 78]. Total number of fixations is thought to be negatively correlated with the ability to find information [48, 77]–in other words, people look at objects longer when they are working to process information.

While prior work has shown the important of various fields in how people assess social media profiles [8, 25, 30], there has been less focus on how people pay attention to profile fields to assess harassing behavior. The theory of perceptual dehumanization suggests that people process faces of norm violators differently than typical faces, and that atypical face-processing facilitates the support of punishment [34]. Perceptual dehumanization also indicates that social information causes changes in responses. Similarly, the culpable control model suggests that people make a spontaneous evaluation of an offense which leads to a desire to blame the offender [3]. Harassment is a kind of deviation from social norms, which can then require more information to make sense of [23, 41]. Thus, our first hypothesis is that when viewing harassing profiles, people require more time to extract and process the information in that profile.

H1: Observers spend more time on harassing profiles than non-harassing profiles.

Harassment is likely to increase arousal, remaining in memory longer than low-arousal information [19, 32]. Memory can be measured via recall, or a person's ability to draw upon a piece of

information from memory and report on it. The progression from attention to memory signals that encoding and storage has taken place place [60]. Memory is a crucial part of social and cognitive experiences—people rely on the ability to remember the past in their lived experiences of the world. Memory can be measured via recall, or a person's ability to draw upon a piece of information from memory and report on it [38, 77, 78]. Thus, we examine whether harassing profiles enter memory via a recall task with the following hypotheses:

H2a: Observers recall harassing profiles in greater detail than non-harassing profiles.

H2a: Observers require less time to recall harassing profiles than non-harassing profiles.

Research shows that people make inferences about others, often based on first impressions [5]. Interactions with a stranger facilitate opportunities for assessment, typically associated with visual perception (of face, body type, clothing choices, posture, etc.) [5]. However, social media profiles offer limited opportunities for assessment. Cues are sparse online [75] and people typically have only one source from which to assess information: the site on the screen in front of them. Within that screen, users may rely on a range of features to make those assessments such as profile photos, prior post history, and social interactions [30, 79]. Information seeking theory shows that people actively and passively process channels of information to make sense of what they see [74]. On Twitter, users rely on content [6] as well as heuristics from usernames when making credibility assessments [57]. Prior studies of website design as well as in marketing and advertising show that people tend to pay attention to images more than text, and to fewer messages more than a cluttered design [62, 78, 79]. However, these prior studies have not explored attention or processing of profile elements in the context of online harassment. To explore this, we ask the following research question:

RQ1: Which profile elements are more salient when processing harassing versus non-harassing profiles?

We examine these hypotheses and research question via an experiment that measured participant's behaviors after engaging with harassing content online. Our experiment was designed to focus on what content participants looked at on profile pages. We chose a lab environment to prioritize controlling interface designs and conditions. In the next section we describe our experiment design and data analysis, interviews and analysis, and participant demographics.

2 METHODS

We conducted an in-lab experiment using a Tobii X2-30 eye-tracker. Participants were consented then participated in the main experiment which involved viewing a series of social media profiles while being eye-tracked. Presentation of stimuli and recording of participants' decisions were conducted using MediaLab [1]. After the experiment, participants completed a brief recall task and an interview study. We compensated participants $25. The study took about one hour to complete. Both the pre-test and the main study were approved by the research team's Institutional Review Board.

2.1 Experiment Design

Participants were instructed that we were interested in learning about what kinds of people they would be interested in following on social media. This prompt was designed to encourage them to evaluate the profiles while remaining unaware of the experimental manipulation. The study consisted of three different visual components.

The first component was the set of primes which consisted of post + reply pairs between two social media users. Each prime contained counterbalanced harassing and non-harassing replies to a post (see Figure 1). The reply was part of the experimental manipulation and varied between a harassing reply and a non-harassing reply. Using an approach from Blackwell, et al. [13], the mock

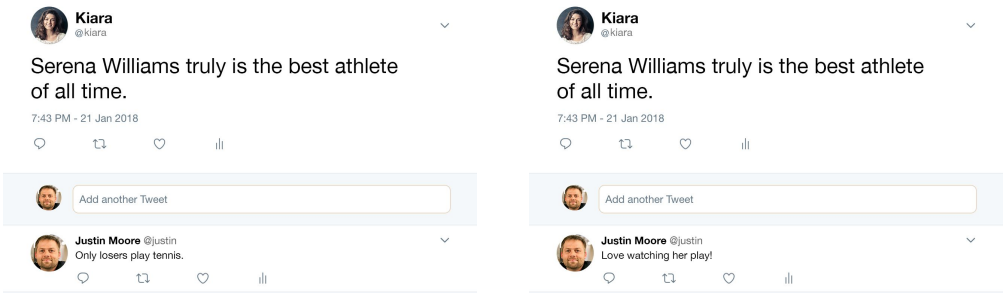

Fig. 1. Example prime for harassing and non-harassing conditions.

profiles used masculine-presenting gender profiles for all harassing and non-harassing conditions. This choice allowed us to keep the study design to a reasonable scope, but prevented us from being able to analyze how gender, or other characteristics like race, might impact results. We used different text and profile photos for each post + reply pair so that participants never saw the same content twice. To develop the primes, we iteratively brainstormed and discussed harassing replies. After we developed a list of possible replies, we evaluated them for consistency. Specifically, we pre-tested the harassing prime text on Mechanical Turk for perceptions of inappropriateness, emotion, and arousal. We asked workers to rate the primes on how appropriate they are using a 7-point Likert scale. We measured emotion and arousal using the using the Self-Assessment Manikin (SAM) [15]. We ran multiple chi-square analyses to compare Likert-type scale responses across all variables and measures. Any prime that was significantly different from the others was revised and then we retested the entire set of stimuli again. There were 64 participants in the first round and 63 new participants in the second round. Median compensation was around $11/hour; the first few participants in both batches completion time was faster and met our target of $15/hour but subsequent participants were slower which meant lower payment (we are taking this into account in future study designs).

The second component was the stimulus, which consisted of social media user profiles for Facebook, Instagram, and Twitter (see Figure 2a, b, c). The mock profiles were built with HTML and CSS using the original dimensions, including height, width, colors, and fonts, of the respective social media platform. In case of proprietary fonts and other content, open-source resources that closely resemble the original content were used. Profiles were also designed to be broadly consistent with the culture of the platform. Profiles were created for each user in the reply part of the prime (e.g. Nathan Smith in Figure 1).

Finally, the third component consisted of the recall task. After finishing six main trials, the software displayed six recall pages corresponding to the six profiles in random order. Participants were instructed to type anything they remembered about the user in a textbox.

We conducted an experiment where participants were exposed to 10 prime + stimuli + recall conditions. The first four prime + stimuli pairs (YouTube, Yelp) were used as warmups to familiarize participants and reduce novelty effects. After the four warm-ups, six main prime + stimuli pairs (Facebook, Twitter, Instagram) were displayed in randomized order. We programmed the MediaLab software to show two of each site (e.g. 2 YouTube conditions, 2 Yelp conditions, 2 Facebook conditions, etc). It also randomized whether participants saw the harassing reply (experiment) or non-harassing reply (control) for each of the 10 conditions, while ensuring that each participant saw a total of 5 of each (5 harassing, 5 non-harassing). Participants then completed a recall task of the profiles in the 6 conditions.

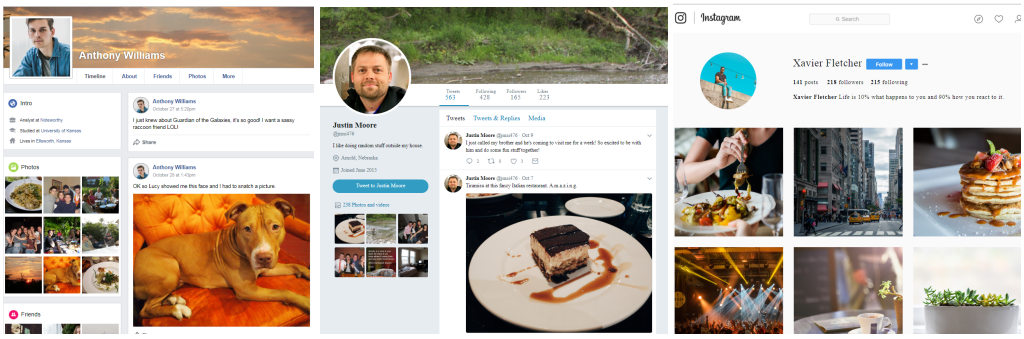

Fig. 2. Facebook, Twitter, and Instagram profile mockups (second versions of each not shown).

We used the following eye movement metrics generated from Tobii Studio: time to first fixation, total visit duration, and fixation count. Prior work suggests that the number of fixations, not their duration, is related to the amount of information a consumer extracts from an ad [77]. We measure both in this study. For hypothesis testing, we used the following metrics: 1) time to first fixation to test saliency of each area of interest (AOI); 2) total visit duration to test how much visual attention participants spent on the AOI; and; 3) the number of fixations they made on each AOI. To prepare eye tracking data for analysis, we defined AOIs in Tobii Studio. For profile pages, AOIs of interest include profile photos, cover photo, full name, profile information, friends, and post history/content. To test the relationship between harassment and eye movements, we used R and the lme4 package to perform linear mixed effects analyses with harassing/non-harassing as the fixed effect and individual differences as the random effect.

We measured whether participants were paying attention to certain information or not by their eye movements. We entered eye movement metrics generated by Tobii Studio into the model as dependent variables, harassment as a fixed effect, and intercepts for participants and social network sites as random effects. We log-transformed continuous parameters, such as total fixation duration, total visit duration, and time to first fixation. Count parameters, such as fixation counts and fixations before first gaze, were fit with generalized linear mixed models with the same fixed and random effects as above.

2.2 Interview

We asked participants to reflect on what parts of a social media profile they use to evaluate people and about their experiences and attitudes related to online harassment. We also showed participants their eye movement on the Tobii software for each of the six stimuli from the experiment as an object for reflection (see Figure 3).

Interviews lasted 28 minutes on average and ranged from 17-47 minutes. We audio recorded and transcribed interviews using the service Rev.com. One member of the research team coded one interview transcript in ATLAS.ti using a preliminary codebook, then the team met to discuss and refine the codebook. The codebook contained 31 codes focused on how people evaluate other people online. Two members of the research team then coded one interview transcript independently. The resulting interrater reliability between coders was Krippendorff's alpha=0.863 which is generally deemed acceptable [50] and they coded the remaining transcripts.

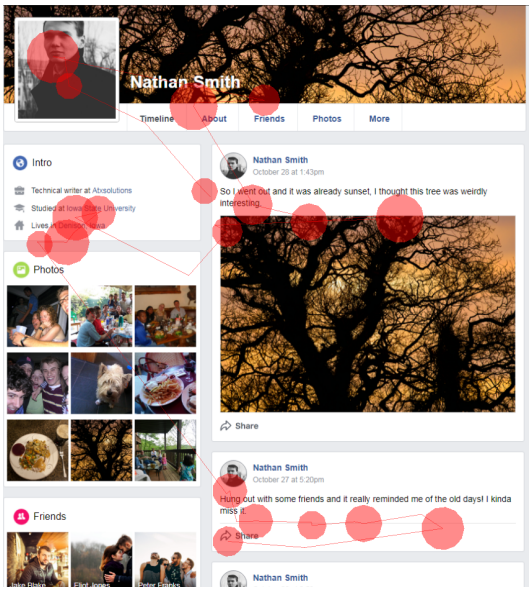

Fig. 3. Screenshot of attention to a profile (P35). Participants were shown the video replay.

2.3 Recruitment and Demographics

We recruited participants from university mailing lists and from outside the university committee using community mailing lists and physical flyers. Recruitment criteria included being able to come to our lab, being 18 years old or older, using the Internet once a week or more, and being able to sit at a typical computer for up to 15 minutes. People with significant vision impairments (e.g., cannot see a dot on a screen) were not eligible to participate.

The age of participants ranged from 19-67 (mean: 30; median: 25). Participants identified as female (N=30), male (N=13), queer/femme (N=1), and one preferred not to answer. 27 were liberal or very liberal, 13 were moderate, and 5 were conservative or very conservative. The sample skewed towards educated (19 had some graduate school or a Masters or professional degree) but also had a significant minority who were not highly educated (12 with some high school or a high school degree and 14 with a bachelor's degree). Of the 45 participants, 34 were employed either full-time or part-time and 20 were students. The remaining were retired (N=3) or unable to work (N=1). Participants identified as White (N=30), Asian (N=9), Black or African American (n-4), Hispanic, Latino/a/x, or Spanish origin (N=4), and Native Hawaiian or Other Pacific Islander (N=1). Per our recruitment criteria, participants were generally active social media users. Facebook was used most actively with 31 participants using it multiple times a day or more. Twitter and Instagram were used multiple times a day or more by 21 and 16 participants, respectively.

3 RESULTS

3.1 Time on harassing profiles

H1 hypothesized that participants would spend more time browsing harassing user profiles. Time was measured by total visit duration and by number of fixations. The total visit duration metric revealed no significant different in time spent on harassing user profiles. Participants made a significantly greater number of fixations on non-harassing profiles than on harassing profiles. ($b = 0.06$, $SE = 0.01$, $p < .001$). They also generated significantly more fixations on non-harassing

post histories than on harassing post histories (b = 0.15, SE = 0.02, p < .001). Other elements including profile information, user name, profile photo, and number of followers/following were not significantly different.

We observed two different types of participant behavior in response to the abusive posts. Many participants explicitly said that they spent less time viewing the post because they did not want to spend time paying attention to someone who was mean, rude, or offensive. For example, P42 said:

"Usually when people do stuff like that, it makes me immediately not interested in them. Like I'm not ... if I weren't doing this, if I saw somebody was posting something negative, I wouldn't click on their profile and I wouldn't look into them. I would just try to get away from them.

In contrast, many others noted that they spent more time trying to make sense of the post. For example, P44 said:

I think the reason I focused so much on his posts was to see if he posted other negative things, or was complaining, or anything. But everything was positive, so that was weird to me, because he's posting all these normal, fun things, but his comment wasn't positive on his friend's post.

Most participants sought context cues when trying to make sense of the abusive comment. They expected coherence in people's social behavior, and turned to profile information and post history to confirm those expectations. Some participants noted that they wanted to see if the abusive comment was in character, or if "he's having a bad day kind of thing" (P3). This was consistent with their responses to negative posts from friends. As P2 noted, if a friend seemed sad, P2 would "go to their page and say, What's going on? Like what's motivating them?"

3.2 Recall of harassing profiles

H2a hypothesized that observers recall harassing profiles in greater detail than non-harassing profiles. Two coders independently coded responses (N=264; 44 participants x 6 profiles) for 1) whether participants recalled the prime and 2) participants' recall of the profile as positive, negative, or neither of those (neutral, can't tell). We calculated agreement between coders using Cohen's κ for both ratings. Prime recall agreement was $\kappa = 0.809$ and profile recall agreement was $\kappa = 0.426$. Because most of the profile recall lack of agreement related to the "neither" option, we focused on those that were rated only positive or negative. We ran two-tailed t-tests to compared profiles rated only as positive or negative. We find no significant difference in prime recall. However, participants are significantly more likely to recall harassing profiles than non-harassing profiles ($p < 0.05$).

H2b predicted that participants require less time to recall harassing profiles than non-harassing profiles. The differences between harassing and non-harassing recall times were not statistically significant. Participants reported trying to reconcile specific inconsistencies from abusive profiles that may have triggered recall. For example, P1 said:

That picture was kinda nice and then he complained about the weather, but he had just told someone else off for complaining about weather, so I thought that was dumb.

Another participant, P36, said that the comment was "a little rude" and they were surprised not to see similarly strong opinions in the profile. P3 recognized that they were more "judge-y" about content on the profile after seeing the abusive comment, which led P3 to feel it was "just some rude guy who's on Twitter." Many of the efforts to find coherence signal confirmation biases among participants. P7 said:

And, I feel like you build up like this resistance toward them, or finding little things that you just don't like about them. So, you just kind of go into depth with all their posts, and everything.

3.3 Salience of profile elements

RQ1 asked which elements on the profile pages are more salient when evaluating harassing profiles. In general, for both conditions, participants' viewing trajectory is to look first at the top of the page

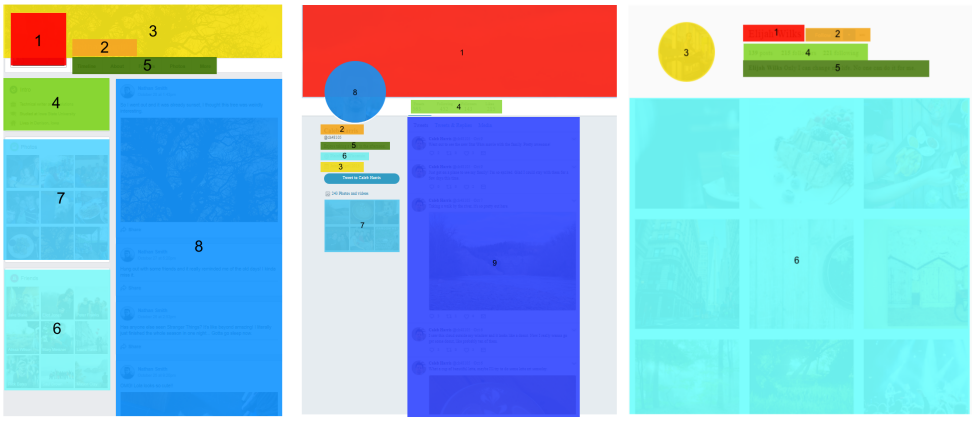

Fig. 4. Facebook, Twitter, and Instagram maps of view order of profile elements for the harassing condition.

and then the left side then the center and right side, containing the post history (see Figure 4). On Facebook, profile elements like photos, information, and cover photo are looked at first but receive little sustained attention. On Twitter, cover photos are similarly unattended to. Both Facebook and Twitter post histories are looked at for longer than Instagram post histories.

We compared the time to first fixation metric for elements of the profile. Participants shifted their attention more quickly to the harassing user's post history compared to non-harassing users ($b = 0.11$, $SE = 0.04$, $p < .05$) (see Figure 5, left). However, they spend more time looking at non-harassing users' post histories than harassing users ($b = 0.11$, $SE = 0.04$, $p < .05$) (see Figure 5, right). Other elements were not significantly different across harassing and non-harassing conditions.

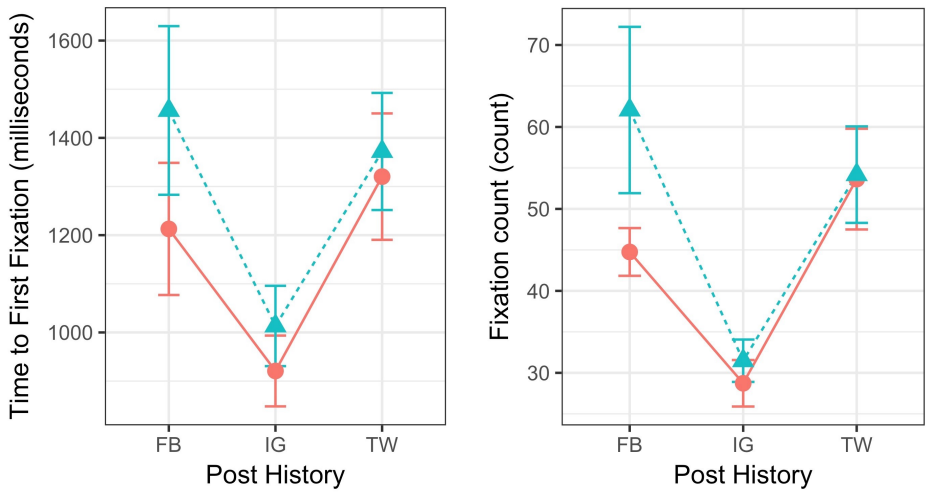

Fig. 5. Time to fixation on post history and fixation count on post history. Blue triangles refer to non-harassing condition and red circles refer to harassing condition.

4 DISCUSSION

This section focuses on two themes. The first examines the idea of post histories as something platforms or users might curate. The second reflects on measurement of attention for understanding social behavior, especially harmful experiences like harassment.

4.1 Curating Post Histories for Assessing Harassers

Results show that participants shift attention to harassing users' post history more quickly than non-harassing users' post history when they are assessing harassing profiles, but there are not differences for other profile elements like name or photo. One explanation for this behavior is that post history might offer a stronger warranting signal [76]—i.e., a signal of authenticity—than other elements. Though prior studies of warranting suggest that comments coming from another person rather than the self provide greater evidence of authenticity [76], the work required to generate a post history might make it similarly difficult to be faked compared to profile elements like bios or photos. Donath's signaling theory explains how assessment signals–those that are costly or difficult to mimic deceptively–are more reliable signals because they cannot be easily manipulated [27]. It is easy to choose a profile photo and bio aligned with one's self-presentational goals, but it takes more work to generate post content and sustain it over time. We cannot tell from our study what users were looking for in post histories; future work could examine how to support post history curation. In some cases, users may want to see most recent posts, but in others, they may wish to see what a user has said about a particular topic or prior responses to a particular user.

Our results also reveal that participants recall the profiles of harassing behaviors more than those of non-harassing profiles, which may indicate greater cognitive processing. This suggests that harassing behavior demands attentional resources, leaving less attention available for other, presumably more desirable, information to be filtered in. One explanation is that people pay more attention to discover context cues that help them understand the behavior of the norm violator [4, 34]. Participants may have expected coherence in people's social behavior, and turned to profile information and post history to confirm or reject those expectations. One caveat with our study design was that it did not add contextual history; participants were surprised to see harassment without any reason to expect it (e.g. from an unknown user in response to a post about the weather). In naturalistic settings online, people may expect harassment based on prior knowledge of the topic, the user who is posting, or the user who is replying and they may process those experiences differently than they do in a lab setting. Still, requiring users to evaluate harassers can be burdensome, especially if they have to do so repeatedly or if doing so is traumatizing for them. In addition to having to report and block each harasser individually, even just seeing them on the feed may be cognitively tasking with cumulative harmful effects over time [43, 54, 73].

4.2 Making Passive Behavior Visible

Many researchers rely on external access to social media data via APIs or scraped data [35] which reveals user behavior in the form of content shared and actions taken (e.g., liking or retweeting content). Industry researchers have a more extensive view into user behavior with access to metrics like time on pages, scroll behavior, friendship links, logins, and other proprietary metrics as well as the ability to pair system behavior with user feedback (e.g. [17, 18]). Across both academic and industry research, system data has allowed researchers to gain deep insights into social behavior online in a variety of domains (e.g. crises [16, 69], political news [10], and mental health [26]). However, these data allow only a partial lens into what people are paying attention to online, relying primarily on content that people post or react to to infer underlying attitudes and states. People may engage with content in meaningful ways, even if they do not click on a post [31].

Indeed, Backstrom et al. note that "although a metric for balance of social attention is potentially useful and theoretically interesting it has been difficult to study empirically" [9]. Studies that seek to measure more nuanced interpretations of attention themselves also rely on crude representations of attention. Nonnecke and Preece relied on self-reports of participation to measure lurking in discussion boards [59]. Using video recordings, Oeldorf-Hirsch et al. measure nonverbal responses to an embarrassing Facebook post and show evidence of a variety of responses like mouth opening, smiling, and taking breaths [61]. Bernstein et al.'s study of perceptions of audience size established a measure based on Facebook news feed item remaining in the active viewport for at least 900 milliseconds [12]. Backstrom et al.'s measures of social attention relied on system logs of profile views, photo views, comments, messages, and wall posts [9].

Studies of collective attention–i.e., large-scale attention to a topic online—similarly rely on system data as proxies for attention. Mitra and Gilbert measure collective attention as the aggregate temporal signatures of an event's reportage (i.e., timestamps on tweets about the event) [56]. Wu and Huberman study news stories posted to dig as measures of large scale collective attention [80]. Our own study presented here also requires estimating attention since people may visually attend to content without deeply processing it or remembering it later. One potential confound, as noted above, is that since profiles did not contain evidence of prior harassing behavior, some of the results could be explained by inconsistency between post and history rather than solely by trying to make sense of the history itself.

4.3 Limitations

While eye-tracking can be a useful technique for more accurately understanding social media use and triangulating findings with other methods, it should be implemented with caution to ethical and interpretative limitations. Though it is unlikely a person could be identified from eye-movements alone, they could be identified if those data were triangulated with other personalized tracking. Eye-tracking also excludes some populations, including people who are Blind and low vision or who have other disabilities that impact their eye movement or attention behaviors.

This work also relied on a lab context where participants likely exhibit heightened attention to profiles, and where they may exhibit various response biases. One way to extend this work to naturalistic environments is to inject short surveys into people's social media experiences after they see abusive content in their news feed to measure recall of and reaction to the content. Alternatively, remote web tracking software is becoming more advanced and could be used in an opt-in online experiment to examine people's naturalistic behaviors while consuming their feeds, though protection of data privacy would be critical. We did not manipulate profile identities, as mentioned above, but it is likely that identity shapes how people experience and observe harassment. Outside of general social media use, we did not ask about social media experiences, such as whether participants had experienced harassment or harassed others. Though these experiences should not affect our experimental design, it could be useful to understand how participants' prior experiences might shape their study behaviors.

5 CONCLUSION

This study examined people's attention to profiles of harassers on Facebook, Instagram, and Twitter. It shows that participants pay attention to harassing users' post history more quickly than to non-harassing users. Participants also tend to recall details about harassing profiles more than non-harassing profiles. However, they do not spend more total time on harassing profiles. This work highlights the importance of social media histories in evaluating others, and proposes that the ability to curate post histories on social media profiles may help users to assess each other more efficiently which is especially important in the case of online harassment.

ACKNOWLEDGMENTS

We thank Tianying Chen for his assistant in designing the profiles and designing and analyzing the pre-test. This material is based upon work supported by the National Science Foundation under Grants 1763297 and 1552503.

REFERENCES

[1] ND. MediaLab and DirectRT Psychology Software. (ND). "http://www.empirisoft.com/medialab.aspx"
[2] Muna Abbas, Elaf Al-Wohaibi, Jonathan Donovan, Emma Hale, Tatyana Marugg, JonB Sykes, Molly K Land, and Richard Ashby Wilson. 2019. Invisible Threats: Online Hate Speech Against Human Rights Defenders in Guatemala. *American Bar Association Center for Human Rights* (2019).
[3] Mark D Alicke. 2000. Culpable control and the psychology of blame. *Psychological bulletin* 126, 4 (2000), 556.
[4] Mark D Alicke, David Rose, and Dori Bloom. 2011. Causation, norm violation, and culpable control. *The Journal of Philosophy* 108, 12 (2011), 670–696.
[5] Nalini Ambady and John Joseph Skowronski. 2008. *First impressions*. Guilford Press.
[6] Paul André, Michael Bernstein, and Kurt Luther. 2012. Who gives a tweet? Evaluating microblog content value. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. 471–474.
[7] Carolina Are. 2021. The Shadowban Cycle: an autoethnography of pole dancing, nudity and censorship on Instagram. *Feminist Media Studies* (2021), 1–18.
[8] Chantal Bacev-Giles and Reeshma Haji. 2017. Online first impressions: Person perception in social media profiles. *Computers in Human Behavior* 75 (2017), 50–57.
[9] Lars Backstrom, Eytan Bakshy, Jon Kleinberg, Thomas Lento, and Itamar Rosenn. 2011. Center of attention: How facebook users allocate attention across friends. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 5. 34–41.
[10] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (2015), 1130–1132.
[11] Lee R Berney and David B Blane. 1997. Collecting retrospective data: accuracy of recall after 50 years judged against historical records. *Social science & medicine* 45, 10 (1997), 1519–1525.
[12] Michael S Bernstein, Eytan Bakshy, Moira Burke, and Brian Karrer. 2013. Quantifying the invisible audience in social networks. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 21–30.
[13] Lindsay Blackwell, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. 2018. When online harassment is perceived as justified. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 12.
[14] Lindsay Blackwell, Jill Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and its consequences for online harassment: Design insights from heartmob. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–19.
[15] Margaret M Bradley and Peter J Lang. 1994. Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry* 25, 1 (1994), 49–59.
[16] Cody L Buntain and Jung Kyu Rhys Lim. 2018. # pray4victims: Consistencies in response to disaster on Twitter. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–18.
[17] Moira Burke and Robert Kraut. 2013. Using Facebook after losing a job: Differential benefits of strong and weak ties. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 1419–1430.
[18] Moira Burke, Cameron Marlow, and Thomas Lento. 2010. Social network activity and social well-being. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1909–1912.
[19] Larry Cahill and James L McGaugh. 1998. Mechanisms of emotional arousal and lasting declarative memory. *Trends in neurosciences* 21, 7 (1998), 294–299.
[20] Robyn Caplan and Tarleton Gillespie. 2020. Tiered governance and demonetization: The shifting terms of labor and compensation in the platform economy. *Social Media+ Society* 6, 2 (2020), 2056305120936636.
[21] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internet's hidden rules: An empirical study of Reddit norm violations at micro, meso, and macro scales. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–25.
[22] Eshwar Chandrasekharan, Mattia Samory, Anirudh Srinivasan, and Eric Gilbert. 2017. The bag of communities: Identifying abusive behavior online with preexisting internet data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 3175–3187.
[23] Robert B Cialdini and Melanie R Trost. 1998. Social influence: Social norms, conformity and compliance. (1998).
[24] Danielle Keats Citron. 2017. Extremist speech, compelled conformity, and censorship creep. *Notre Dame L. Rev.* 93 (2017), 1035.

[25] Scott Counts and Kristie Fisher. 2011. Taking it all in? visual attention in microblog consumption. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 5. 97–104.

[26] Munmun De Choudhury, Scott Counts, Eric J Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 626–638.

[27] Judith Donath. 2007. Signals in social supernets. *Journal of Computer-Mediated Communication* 13, 1 (2007), 231–251.

[28] Evelyn Douek. 2021. Governing online speech: From" posts-as-trumps" to proportionality and probability. *Colum. L. Rev.* 121 (2021), 759.

[29] Maeve Duggan. 2017. Online harassment 2017. (2017).

[30] Nicole B Ellison, Jeffrey T Hancock, and Catalina L Toma. 2012. Profile as promise: A framework for conceptualizing veracity in online dating self-presentations. *New media & society* 14, 1 (2012), 45–62.

[31] Nicole B Ellison, Penny Triu, Sarita Schoenebeck, Robin Brewer, and Aarti Israni. 2020. Why we don't click: Interrogating the relationship between viewing and clicking in social media contexts by exploring the "non-click". *Journal of Computer-Mediated Communication* 25, 6 (2020), 402–426.

[32] Michael Eysenck. 2012. *Attention and arousal: Cognition and performance.* Springer Science & Business Media.

[33] Jordan Fairbairn. 2020. Before# MeToo: Violence against women social media work, bystander intervention, and social change. *Societies* 10, 3 (2020), 51.

[34] Katrina M Fincher and Philip E Tetlock. 2016. Perceptual dehumanization of faces is activated by norm violations and facilitates norm enforcement. *Journal of Experimental Psychology: General* 145, 2 (2016), 131.

[35] Deen Freelon. 2018. Computational research in the post-API age. *Political Communication* 35, 4 (2018), 665–668.

[36] Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.

[37] Tarleton Gillespie. 2022. Do Not Recommend? Reduction as a Form of Content Moderation. *Social Media+ Society* 8, 3 (2022), 20563051221117552.

[38] Elizabeth L Glisky, Susan R Rubin, and Patrick SR Davidson. 2001. Source memory in older adults: An encoding or retrieval problem? *Journal of experimental psychology: learning, memory, and cognition* 27, 5 (2001), 1131.

[39] Carrie Goldberg. 2019. *Nobody's Victim: Fighting Psychos, Stalkers, Pervs, and Trolls.* Penguin.

[40] Eric Goldman. 2021. Content Moderation Remedies. *Mich. Tech. L. Rev.* 28 (2021), 1.

[41] Erich Goode and Nachman Ben-Yehuda. 2010. *Moral panics: The social construction of deviance.* John Wiley & Sons.

[42] Kishonna L Gray. 2012. Intersecting oppressions and online communities: Examining the experiences of women of color in Xbox Live. *Information, Communication & Society* 15, 3 (2012), 411–428.

[43] Kishonna L Gray. 2020. Black Gamers' Resistance. *Race and Media: Critical Approaches* (2020), 241–51.

[44] Oliver L Haimson, Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–35.

[45] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138* (2017).

[46] Shagun Jhaver, Amy Bruckman, and Eric Gilbert. 2019. Does transparency in moderation really matter? User behavior after content removal explanations on reddit. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–27.

[47] Jialun Aaron Jiang, Morgan Klaus Scheuerman, Casey Fiesler, and Jed R Brubaker. 2021. Understanding international perceptions of the severity of harmful content online. *PloS one* 16, 8 (2021), e0256762.

[48] M Just and P Carpenter. 1976. Eye fixations and cognitive processes. Cognitive Psychology. (1976).

[49] Kate Klonick. 2017. The new governors: The people, rules, and processes governing online speech. *Harv. L. Rev.* 131 (2017), 1598.

[50] Klaus Krippendorff. 2004. Reliability in content analysis: Some common misconceptions and recommendations. *Human communication research* 30, 3 (2004), 411–433.

[51] Amanda Lenhart, Michele Ybarra, Kathryn Zickuhr, and Myeshia Price-Feeney. 2016. *Online harassment, digital abuse, and cyberstalking in America.* Data and Society Research Institute.

[52] Brandeis Marshall. 2021. Algorithmic misogynoir in content moderation practice. *Heinrich-Böll-Stiftung European Union* (2021).

[53] Alice E Marwick. 2021. Morally motivated networked harassment as normative reinforcement. *Social Media+ Society* 7, 2 (2021), 20563051211021378.

[54] Adrienne Massanari. 2017. # Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New media & society* 19, 3 (2017), 329–346.

[55] J Nathan Matias. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proceedings of the National Academy of Sciences* 116, 20 (2019), 9785–9789.

[56] Tanushree Mitra, Graham Wright, and Eric Gilbert. 2017. Credibility and the dynamics of collective attention. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–17.
[57] Meredith Ringel Morris, Scott Counts, Asta Roseway, Aaron Hoff, and Julia Schwarz. 2012. Tweeting is believing? Understanding microblog credibility perceptions. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*. 441–450.
[58] Odmar Neumann. 1996. Theories of attention. In *Handbook of perception and action*. Vol. 3. Elsevier, 389–446.
[59] Blair Nonnecke and Jenny Preece. 2000. Lurker demographics: Counting the silent. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 73–80.
[60] Donald A Norman. 1976. Memory and attention: An introduction to human information processing. (1976).
[61] Anne Oeldorf-Hirsch, Jeremy Birnholtz, and Jeffrey T Hancock. 2017. Your post is embarrassing me: Face threats, identity, and the audience on Facebook. *Computers in human behavior* 73 (2017), 92–99.
[62] Katharina Reinecke, Tom Yeh, Luke Miratrix, Rahmatri Mardiko, Yuechen Zhao, Jenny Liu, and Krzysztof Z Gajos. 2013. Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2049–2058.
[63] Sarah T Roberts. 2019. *Behind the screen*. Yale University Press.
[64] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R Brubaker. 2021. A framework of severity for harmful content online. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–33.
[65] Sarita Schoenebeck and Lindsay Blackwell. 2020. Reimagining social media governance: Harm, accountability, and repair. *Yale JL & Tech.* 23 (2020), 113.
[66] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. 2021. Drawing from justice theories to support targets of online harassment. *new media & society* 23, 5 (2021), 1278–1300.
[67] Norbert Schwarz. 2007. Retrospective and concurrent self-reports: The rationale for real-time data capture. *The science of real-time data capture: Self-reports in health research* 11 (2007), 26.
[68] Joseph Seering, Tony Wang, Jina Yoon, and Geoff Kaufman. 2019. Moderator engagement and community development in the age of algorithms. *New Media & Society* 21, 7 (2019), 1417–1443.
[69] Kate Starbird, Jim Maddock, Mania Orand, Peg Achterman, and Robert M Mason. 2014. Rumors, false flags, and digital vigilantes: Misinformation on twitter after the 2013 boston marathon bombing. *IConference 2014 proceedings* (2014).
[70] Sharifa Sultana, Mitrasree Deb, Ananya Bhattacharjee, Shaid Hasan, SM Raihanul Alam, Trishna Chakraborty, Prianka Roy, Samira Fairuz Ahmed, Aparna Moitra, M Ashraful Amin, et al. 2021. 'Unmochon': A Tool to Combat Online Sexual Harassment over Facebook Messenger. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.
[71] Anne M Treisman and Garry Gelade. 1980. A feature-integration theory of attention. *Cognitive psychology* 12, 1 (1980), 97–136.
[72] Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. ICWSM—a great catchy name: Semi-supervised recognition of sarcastic sentences in online product reviews. In *fourth international AAAI conference on weblogs and social media*.
[73] Nikki Usher, Jesse Holcomb, and Justin Littman. 2018. Twitter makes it worse: Political journalists, gendered echo chambers, and the amplification of gender bias. *The international journal of press/politics* 23, 3 (2018), 324–344.
[74] Joseph B Walther and Judee K Burgoon. 1992. Relational communication in computer-mediated interaction. *Human communication research* 19, 1 (1992), 50–88.
[75] Joseph B Walther and Malcolm R Parks. 2002. Cues filtered out, cues filtered in: Computer-mediated communication and relationships. *Handbook of interpersonal communication* 3 (2002), 529–563.
[76] Joseph B Walther, Brandon Van Der Heide, Lauren M Hamel, and Hillary C Shulman. 2009. Self-generated versus other-generated statements and impressions in computer-mediated communication: A test of warranting theory using Facebook. *Communication research* 36, 2 (2009), 229–253.
[77] Michel Wedel and Rik Pieters. 2000. Eye fixations on advertisements and memory for brands: A model and findings. *Marketing science* 19, 4 (2000), 297–312.
[78] Michel Wedel, Rik Pieters, et al. 2008. Eye tracking for visual marketing. *Foundations and Trends® in Marketing* 1, 4 (2008), 231–320.
[79] Max Weisbuch, Zorana Ivcevic, and Nalini Ambady. 2009. On being liked on the web and in the "real world": Consistency in first impressions across personal webpages and spontaneous behavior. *Journal of Experimental Social Psychology* 45, 3 (2009), 573–576.
[80] Fang Wu and Bernardo A Huberman. 2007. Novelty and collective attention. *Proceedings of the National Academy of Sciences* 104, 45 (2007), 17599–17601.

Received May 2022; revised August 2022; accepted September 2022