Reddit Insights: Improving Online Discussion Culture by Contextualizing User Profiles

Franz Waltenberger*
Center for Digital Technology and Management, Germany
waltenberger@cdtm.de

Simon Höferlin†
Center for Digital Technology and Management, Germany
simon.hoeferlin@cdtm.de

Michael Froehlich‡
Center for Digital Technology and Management, Germany
froehlich@cdtm.de

ABSTRACT

More then forty years after the creation of the first online messaging board, the quality of online discussion culture remains a topic of significant scientific and public debate. Social media platforms have struggled to maintain a free marketplace of ideas while addressing issues such as censorship, propaganda, and misinformation. Traditional methods of content moderation have been ineffective in addressing these issues with incivility in online discussions remaining a major concern. In an effort to address the ongoing challenges facing online discussion culture, we have developed a browser extension called Reddit Insights, designed to improve discourse quality through design interventions. The extension provides context to user profiles by aggregating information from past commenting behavior. Results from a small-scale qualitative assessment with 9 users indicate that the extension helped users contextualize posts, identify implicit political tendencies and decide with whom to interact and choose in which discussions to engage more wisely.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Interactive systems and tools**; **Collaborative and social computing systems and tools**.

KEYWORDS

online discussion culture, reddit, computational social science

ACM Reference Format:
Franz Waltenberger, Simon Höferlin, and Michael Froehlich. 2023. Reddit Insights: Improving Online Discussion Culture by Contextualizing User Profiles. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3544549.3585671

1 INTRODUCTION

More then forty years after the creation of the first online messaging board, the quality of online discussion culture remains a topic of significant scientific and public debate [5]. Despite efforts of major social media platforms such as Facebook and Twitter, maintaining a free marketplace of ideas while addressing issues such as censorship, propaganda, and misinformation has shown itself to be a challenging task. Traditional methods of content moderation have proven ineffective in keeping this delicate balance. Politically charged environments exacerbate this problem, as one side may perceive themselves as victims of unfair moderation policies and user agreements. Social media sites have become gatekeepers of information, and even soft moderation techniques, such as reducing the reach of posts classified as misinformation, can have significant effects [21]. Furthermore, the inherent lack of transparency in moderation decisions has led to an uncanny feeling for many users that the content they see has already been approved by a third party and deemed appropriate for them [23].

In addition to issues related to informational content, the general incivility of many online discussions is another major concern for content moderation [16]. Traditional content moderation methods primarily focus on addressing rude behavior after it has occurred, while major platforms such as Facebook, Twitter, and Reddit have yet to implement design features that encourage civil behavior before the fact. One exception from this is YouTube, which recently has launched a feature called "Profile Cards", which quickly allows users to see another commentator's most recent comments as well as their channel subscriptions [13]. While all of the aforementioned platforms emphasize the importance of civil discussion in their user agreements, these guidelines are only rarely read by the average user [14].

To address the ongoing challenges facing online discussion culture, we have developed Reddit Insights, a browser extension designed to improve user engagement and behavior. The extension provides context to Reddit user profiles by aggregating information from past posting behavior. When a Reddit Insights user encounters another user on Reddit, the extension automatically displays the subreddits the other user is most active in and visually indicates if the other poster has a history of writing rude or discourteous posts.

Reddit Insights adds to existing HCI research by offering a novel design approach relying on past user behavior for contextualization. Compared to previous contributions, we expand the analysis by focusing on social signals beyond toxicity and providing participants with detailed social context about their counterparts [9]. Additionally, by focusing on Reddit instead of Twitter, we can examine the effects on a finer scale as discussions are generally longer and allow for more deliberation. While the plugin is currently only working on Reddit, our architecture shown is scalable and allows for extension. The concept could thus be modified for future research on different platforms.

*Also with Technical University of Munich.
†Also with Technical University of Munich.
‡Also with LMU Munich, University of the Bundeswehr Munich.

CHI EA '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3585671

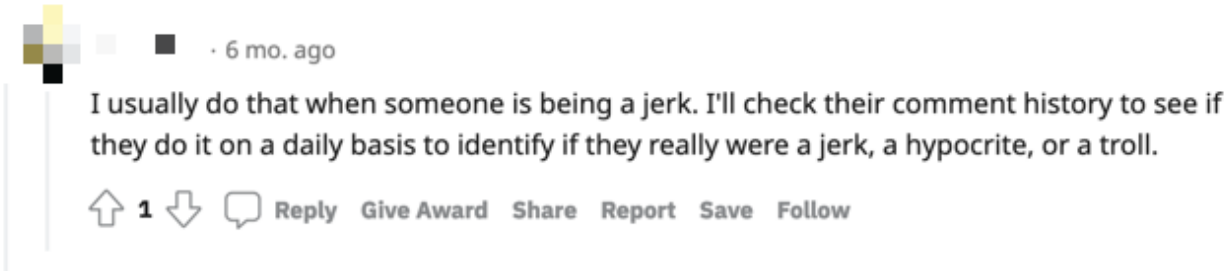

Figure 1: Anecdotal Example of a Reddit User Checking Other Users' Comment History to Identify Uncivil Behavior [12]

In this late-breaking work, we present the results of a small-scale qualitative assessment of the prototype with 9 users. Study participants reported that the extension helped them contextualize posts, identify implicit political tendencies, and distinguish between edgy humor and problematic mindsets. Users furthermore reported that the extension helped them decide with whom to interact and choose in which discussions to engage more wisely. They also reported a perceived better understanding of other users' credibility when discussing more complex topics requiring in-depth knowledge and found it easier to cope with aggressive behavior after discovering similar behavior by the same user in the past.

While these results require further analysis through a quantitative study, the qualitative results are promising. By implementing the design features shown here into individual websites, many of the core content moderation issues social media sites are currently facing could be mitigated, in some cases even before they arise.

2 CONCEPT AND PROTOTYPE

Our concept builds on social signaling theory [20]. According to social signaling theory, people use cues such as education, job experience, or status symbols to signal that they are reliable cooperation partners in situations of information asymmetry [20]. Similar cues also exist in the online space such as the number of friends an account has [22]. Users can utilize these cues as indicators about another user's past behavior, which might influence their opinion about them [24] and consequently their willingness to trust and interact with them.

Our approach aims to generate valuable social signals by analyzing data of past user behavior and aggregating it into useful and easy-to-understand interface cues. This approach is similar to existing user behavior on Reddit, as the anecdotal example provided in Figure 1 shows, but automates an otherwise tedious and manual process.

With the the generated cues, we aim to improve the social signals available to people in online discussion. We decided to use the most frequented subreddits, civility of language, and general account information as fitting signals. These cues rely only on data that is readily available for any operator of online discussion forums and social network platforms and thus ensures a broad potential application of the presented approach. An additional advantage is that these cues can be generated without the need for users to provide additional information about them (e.g. profile pictures) and can thus be implemented in a privacy preserving way.

2.1 User Interface

With the installed browser extension, every post and comment on Reddit is augmented by a series of different tags, generated based on the respective users past behavior on the platform. Figure 2 shows the interface injected by the browser extension and the examples of different types of tags.

User profiles are contextualized with three different types of tags:

(1) Behavioral tags display an analysis of users' past behavior. We generate this information by analyzing a user's 250 most recent posts and comments with the machine learning based Perspective API to detect anti-normative behaviors such as toxicity, insults, or identity attacks [2]. Each behavioral tag displays an associated score and respective color-coding for easy identification. For example, if a comment has a toxicity score of 80%, then 8 out of 10 people would argue that a comment contains "*rude, disrespectful, or inappropriate*" parts and "i*s likely to make people leave a discussion*" [3]. Participants were able to click on a user's behavioral tags to see the individual comment responsible for the rating.

(2) Interest tags contextualize users interest by displaying their most active subreddits and how often they contributed to them within their last 250 comments and submissions.

(3) Activity tags provide information about a users platform contributions. They show their overall number of posts and comments as well as their mean rating, calculated from upvotes and downvotes.

At the beginning of the study, each participant received a detailed briefing on how to use the extension. Throughout the study, participants could view detailed explanations of the individual scores by clicking on the extension icon in their browser's address bar.

2.2 Implementation

We implemented the proposed system as a browser extension for Chrome. We selected Reddit as empirical contextas it is the 9th largest website by traffic [19], offers a plethora of different topic-specific subreddits, and has an easily accessible API. Due to the non-availability of browser extensions for both Android and iOS, our prototype was built as a desktop browser extension. This allows for low entry barrier, seamless integration and has been a popular in past HCI research [1, 4, 10].

Figure 3 shows the software architecture of the developed prototype. We decoupled the prototype into 5 subsystems in two environments, namely the user's *Web Browser* and the *Cloud Environment* where

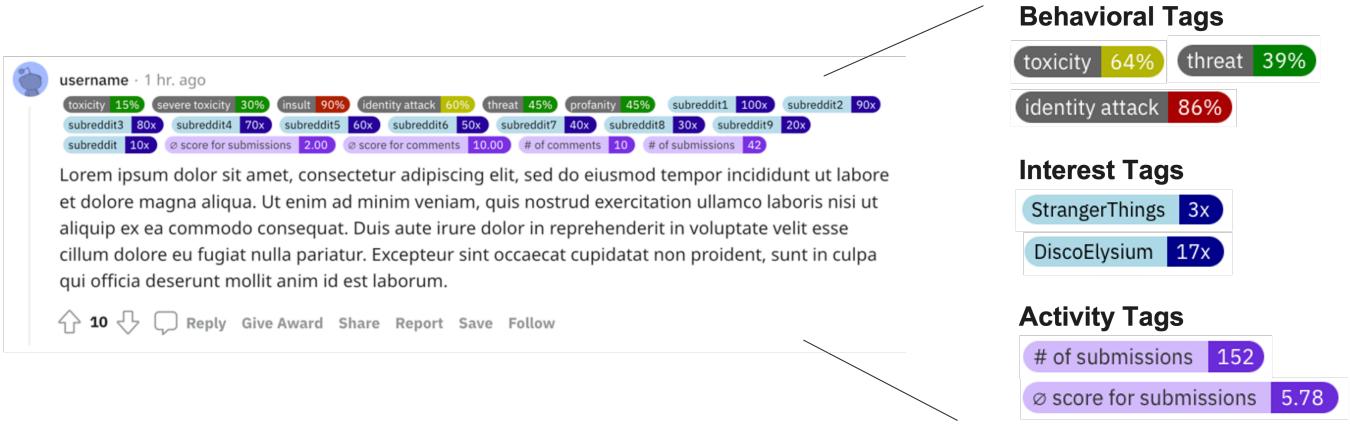

Figure 2: User Interface and Examples of Different Types of Tags

our backend services are deployed. From the user's perspective, the *Browser Extension* adds the described tags to each post and comment on the discussion platform, i.e. Reddit. Whenever a post or comment is viewed, the *Browser Extension* requests the required data via the *API Gateway*, implemented with NodeJS. If data for the respective user is not cached in the *MongoDB* database, the profile and past posts and comments of said user are requested via the *Pushshift* API. The acquired data is then analyzed using *Perspective API* and the results are stored in *MongoDB*. This architecture allows for the possibility to integrate different platforms in the future, while providing fast and scalable load times. In first test we achieved 1-3 second load times for the ad-hoc generation and display of labels.

3 METHOD

We evaluated the developed prototype in a qualitative study (N=9) following a technology probe approach [11]. We distributed the browser extension to 12 frequent users of Reddit. After using the Reddit with the extension we invited them back to semi-structured interviews to elicit qualitative insights about their experiences.

3.1 Sample

Following the 10±2 rule [8] we recruited 12 participants at our university out of which 9 completed the entire study. We made sure to admit only participants who stated that they "regularly" used Reddit. Participants were between 19 and 34 years old. Two (22%) participants identified as female and seven (78%) as male. Those demographics are representative of the Reddit userbase [15].

3.2 Study Design

We asked participants to install the developed browser extension on their desktop devices and use Reddit as they would normally do over a period about two weeks. To maintain users' privacy we only tracked usage time during the study. After completion of the study, we invited participants to a semi-structured interview, conducted online via Zoom, to examine their experiences and collect feedback. All interviews were recorded and transcribed for further analysis. We used thematic analysis, combining deductive and inductive coding, as proposed by Fereday and Muir-Cochrane [6], to analyze the interviews. Theory-driven codes were added based on Friess and Eilders framework for online deliberation [7]. Finally, we asked participants to rate their experience using the User Experience Questionnaire (UEQ) [17, 18].

4 RESULTS

Participants installed and used the browser extension over a period of 11 to 19 days (mean 14.8). Over this period usage time ranged from 15 to 259 minutes (mean 79 minutes). Study participants reported improved contextualization of posts, identification of implicit political tendencies, and that the extension helped them decide with whom to interact and choose in which discussions to engage more wisely. Participants also reported a better understanding of other users' credibility when discussing more complex topics requiring in-depth knowledge and found it easier to cope with aggressive behavior after discovering similar behavior by the same user in the past.

4.1 User Experience and Overall Perception

Overall, the prototype was well accepted. Several participants stated that the intended to continue using the extension. This positive experience is also reflected in the UEQ ratings presented in Table 1. Scores for each UEQ dimension range from -3 to 3, with higher scores relating to a better user experience with regards to the considered dimension. Scores between -0.8 and 0.8 can be interpreted as neutral, scores > 0.8 as positive [18]. Following the benchmark established by Schrepp et al. the scores for the *Attractiveness*, *Perspicuity*, *Stimulation*, and *Novelty* dimensions even rank in the top 25 percentile [18].

4.2 Effects of the Intervention Mechanisms

4.2.1 Less Interaction With Uncivil Users. Several users reported that the contextualization of comments through interest tags changed their own posting behavior. One user reported specifically avoiding interacting with the creator of an uncivil post who wrote an uncivil comment and noted that *"Usually I would have engaged and said*

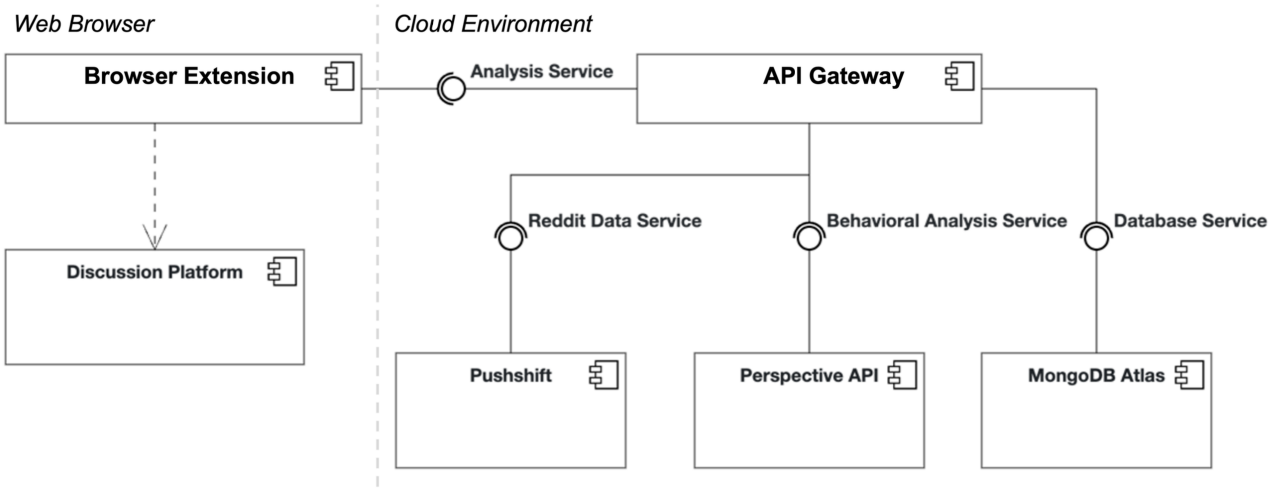

Figure 3: Software Architecture and Subsystems

something, 'Oh, you're an idiot', or something toxic myself. But this person had toxicity and insult flags. So I didn't engage at all. I just reported the comment to the mods and they deleted it." (User 7). The same user noted that they *"started avoiding people that were active [...] in radical or sexist subs"* Another user reported that *"if I see the behavioral tags, and they have bad behavior, I'm not going to start that discussion [...] That definitely changed my behavior and makes me think it made me comment less to be honest."* (User 5).

4.2.2 Improved Credibility Assessment. Participants also stated that in some cases they would ascribe higher or lower credibility to other poster's contributions. One user reported that *"if I see someone posting something in the motorcycle subreddit and I see he's also in another more technical subreddit it might make him more trustworthy"* (User 9). Others reported similar experiences for the topics of personal finance as well as history and politics.

4.2.3 Increased Humanization. Another common theme was that users remained aware that the other users were after all humans on the other side of the keyboard. Users stated they felt like anonymity was decreased and that the extension made them see other commentators more as individuals. One user said that the information gave them *"more a feeling that this is a real person somehow. Maybe you even have some common interest that you wouldn't have thought, but might make people even more empathetic to you."* (User 9). Furthermore users reported that the plugin helped them to rely less on first impressions and make the person behind the post more differentiated than just based on what they wrote.

4.2.4 Increased Self-Reflection. Furthermore, we saw increased self-reflection among users about their own posting habits. While the plugin does not directly display a users' own metrics, by visiting their own profile or just revisiting a comment thread they already interacted with, users can still see their own metrics. One user specifically reflected about how the extension made them *"think twice about what do people actually say. You know, internet language is so different than language you would use everyday. It's so much more informal to describe it nicely, let's say. So it was highlighting that again for me."* (User 4).

Table 1: User Experience Questionnaire (UEQ) Scores (N=9)

Dimension	Score	Benchmark
Attractiveness	1.74	Good (top 25%)
Perspicuity	1.81	Good (top 25%)
Efficiency	1.33	Above Average (top 50%)
Dependability	0.89	Below Average (top 75%)
Stimulation	1.56	Good (top 25%)
Novelty	1.39	Good (top 25%)

4.3 Concerns

Study participants also voiced several concerns regarding the browser extension and the intervention mechanisms utilized within. One concern was that *"there will be more bubbles"* (User 7), with people actively avoiding users from groups they do not identify with. None of the study participants stated that they might actively search for users outside their bubble to challenge their own opinions. Furthermore, one user was concerned that the extension *"could lead to people judging other Reddit users more based on their interest tags, especially. Because sometimes even without the plugin, I can see people going through a Reddit comment or a history of posts before they comment and then judge them based on their history. And if it's so readily accessible [...] it could lead to more people judging your history or your interests."* (User 5).

5 DISCUSSION

In our study we evaluated a browser extension aimed at improving online discussion quality by contextualizing user profiles with data from previous contributions. Users self-reported substantial effects

on their own behavior, including increased recognition of other users as individuals, improved credibility, higher self-reflection and an avoidance of interactions with users expressing themselves in an uncivil way in the past. Study participants generally reported a good user experience, however also criticized the sometimes low dependability of the generated behavioral scores based on Perspective API.

Overall the results show the potential of interface design interventions for civil online discourse. While modern platforms currently mainly rely on ex post content moderation mechanisms, the here presented approach could enable users themselves to make more informed and consequently more satisfying decisions when browsing discussion forums. With anonymity having been previously identified as a key driver for incivility, this kind of intervention could be a privacy-preserving and transparent method combating heightened incivility by increasing the awareness for the human on the other side of the keyboard. Another possible use case for the presented intervention mechanisms could be mediating negative effects on victims of online abuse. Providing contextualization to those affected might help them by either showing general aggressiveness of the perpetrator or help them to avoid potentially abusive users entirely.

The study also shows that descriptive statistics, while sometimes being less informative, also have a key credibility advantage. While Perspective API is a powerful tool to detect incivility, it fails to adapt to text features such as quotes and cannot distinguish reliably between someone quoting an insult and someone actually insulting another user. This unreliable detection lowered the perceived reliability of the tool when users, upon checking the comment responsible for the low score, sometimes found that they did not perceive the comment as offensive themselves. For example, *"What I noticed is that sometimes users would get a toxicity label, but when I looked at the comments, I wouldn't have classified those comments as toxic. Sometimes there were curse words but people were not really targeting other users or something."* (User 9).

While the extension offered a very transparent option to check the classification, users, despite understanding the limitations, put significant faith into the provided scores *"When you click on [toxicity labels], there are posts that they get the data from actually. Sometimes when I click my profile, or other people profile, I see a post and I see no toxicity. So maybe I'm toxic. But I only say that, for example, my friends cooking is really shitty. So that's why they they kind of increase my toxicity score. I don't know if it's really toxic or not. [...] I know my rates are not very high. But even though I see 20 percent I feel like "Oh, my God, why? What did I do?""* (User 6). The results indicate that one needs to be especially careful when using such tools and either make similarly black-boxed classification processes completely transparent or consider cutting them completely for the sake of user trust.

5.1 Limitations and Future Work

This paper presented results from a small qualitative evaluation. The small number of participants (N=9), the short study duration between 11 and 19 days, and data-collection through self-reporting by the users to determine possible intervention effects naturally limit the generalizability of the findings.

While participants reported that they were considerably influenced by the presented cues, future experimental evaluation to test quantitatively whether user behavior is affected by the intervention is needed. With the low reliability of Perspective API for individual online comment classifications, we recommend to not use Perspective API for such use cases in the future but instead either an alternative model or to consider removing behavioral scoring completely, as results indicate that users found the purely descriptive interest and activity tags similarly helpful and more reliable.

Before deploying the application at scale, we also need to discuss ethical considerations of the presented intervention mechanisms. Multiple participants were concerned about the extension being used to judge other users comments' mainly based on their tags. While the extension itself only aggregated publicly available information, the simplicity and availability of it play a significant role. One possible solution to mitigate this issue might be the addition of an additional layer for interested users to go through, for example, by only displaying the information on the click of a button. This could lead to a more targeted application by users only in cases where they would specifically like to decrease the information asymmetry.

Another potential ethical issue of the extension is the implicit promotion of bubble-think and potential general avoidance of users showing undesired traits. Tags generated by the extension simplify making predictions about another users gender, sexuality or political affiliation that otherwise would not have been obvious. By always displaying these traits in unrelated online spaces and not allowing others to hide them when desired, it would be easy for individuals to harass others just based on their related interests.

Additionally, there may be ethical concerns related to the utilization of Perspective API for processing data of other Reddit users. While the strictly descriptive interest and activity tags only show aggregated public data, behavioral tag scores generated by Perspective API represent not a neutral value but one inferred by the underlying model. Potential ways to handle this issue in the future include only analyzing accounts of users that are using the extension themselves, more clearly notifying participants of the underlying bias or dropping such non-descriptive behavioral tags completely.

It should also be noted that none of the study participants stated that they were actively searching for users with cues indicating opposing views in order to challenge their own beliefs. The question of whether this is representative or just due to the small sample size is another promising avenue for future work.

5.2 Conclusion

Due to the combination of Reddit as platform, its scaleability and overall attractiveness to users, the browser extension presented herein opens up many paths for innovative HCI research. All previously theorized effects were preliminary confirmed within the scope of the qualitative study, with the tool design allowing for a large scale quantitative verification of the results in the future. Furthermore, this approach could include exploring more complex scenarios that present users with more varied and detailed information.

REFERENCES

[1] Iñigo Aldalur, Alain Perez, and Felix Larrinaga. 2021. MAWA: A Browser Extension for Mobile Web Augmentation. In *IFIP Conference on Human-Computer Interaction*. Springer, 221–242.

[2] Perspective API. 2022. *Attributes & Languages.* Retrieved Jan 16, 2023 from https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages

[3] Perspective API. 2022. *Limits & Errors.* Retrieved Jan 16, 2023 from https://developers.perspectiveapi.com/s/about-the-api-limits-and-errors

[4] Paul Boyle and Lynsay A Shepherd. 2021. Mailtrout: a machine learning browser extension for detecting phishing emails. In *34th British HCI Conference 34*. 104–115.

[5] Simeon Edosomwan, Sitalaskshmi Kalangot Prakasan, Doriane Kouame, Jonelle Watson, and Tom Seymour. 2011. The history of social media and its impact on business. *Journal of Applied Management and entrepreneurship* 16, 3 (2011), 79.

[6] Jennifer Fereday and Eimear Muir-Cochrane. 2006. Demonstrating Rigor Using Thematic Analysis: A Hybrid Approach of Inductive and Deductive Coding and Theme Development. *International Journal of Qualitative Methods* 5, 1 (2006), 80–92. https://doi.org/10.1177/160940690600500107

[7] Dennis Friess and Christiane Eilders. 2015. A Systematic Review of Online Deliberation Research. *Policy & Internet* 7, 3 (2015), 319–339. https://doi.org/10.1002/poi3.95

[8] Wonil Hwang and Gavriel Salvendy. 2010. Number of People Required for Usability Evaluation: The 10±2 Rule. *Commun. ACM* 53, 5 (may 2010), 130–133. https://doi.org/10.1145/1735223.1735255

[9] Jane Im, Sonali Tandon, Eshwar Chandrasekharan, Taylor Denby, and Eric Gilbert. 2020. Synthesized social signals: Computationally-derived social signals from account histories. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[10] Christina Katsini, Yasmeen Abdrabou, George E Raptis, Mohamed Khamis, and Florian Alt. 2020. The role of eye gaze in security and privacy applications: Survey and future HCI research directions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–21.

[11] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2017. *Research methods in human-computer interaction.* Morgan Kaufmann.

[12] nicokokun. 2021. *R/antimeme - terrifying.* https://www.reddit.com/r/antimeme/comments/w1zaeg/terrifying

[13] Sarah Perez. 2020. *YouTube launches profile cards that show a user's comment history.* https://techcrunch.com/2020/01/15/youtube-launches-profile-cards-that-show-a-users-comment-history/

[14] Victoria C Plaut and Robert P Bartlett. 2011. Blind consent? A social psychological investigation of non-readership of click-through agreements. *Law and human behavior* (2011), 1–23.

[15] Nicholas Proferes, Naiyan Jones, Sarah Gilbert, Casey Fiesler, and Michael Zimmer. 2021. Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics. *Social Media + Society* 7, 2 (2021), 20563051211019004. https://doi.org/10.1177/20563051211019004 arXiv:https://doi.org/10.1177/20563051211019004

[16] Ludovic Rheault, Erica Rayment, and Andreea Musulan. 2019. Politicians in the line of fire: Incivility and the treatment of women on social media. *Research & Politics* 6, 1 (2019), 2053168018816228.

[17] Martin Schrepp, Andreas Hinderks, and Joerg Thomaschewski. 2014. Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios. In *Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience*, Aaron Marcus (Ed.). Springer International Publishing, Cham, 383–392.

[18] Martin Schrepp, Joerg Thomaschewski, and Andreas Hinderks. 2017. Construction of a benchmark for the user experience questionnaire (UEQ). (2017). https://doi.org/10.9781/ijimai.2017.445

[19] Semrush. 2022. *reddit.com – November 2022 Traffic Stats.* Retrieved Jan 16, 2023 from https://www.semrush.com/website/reddit.com/overview/

[20] Michael Spence. 1978. Job market signaling. In *Uncertainty in economics*. Elsevier, 281–306.

[21] Héloïse Théro and Emmanuel M Vincent. 2022. Investigating Facebook's interventions against accounts that repeatedly share misinformation. *Information Processing & Management* 59, 2 (2022), 102804.

[22] Stephanie Tom Tong, Brandon Van Der Heide, Lindsey Langwell, and Joseph B Walther. 2008. Too much of a good thing? The relationship between number of friends and interpersonal impressions on Facebook. *Journal of computer-mediated communication* 13, 3 (2008), 531–549.

[23] Kristen Vaccaro, Christian Sandvig, and Karrie Karahalios. 2020. " At the End of the Day Facebook Does What ItWants" How Users Experience Contesting Algorithmic Content Moderation. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–22.

[24] Brandon Van Der Heide and Young-shin Lim. 2016. On the conditional cueing of credibility heuristics: The case of online influence. *Communication Research* 43, 5 (2016), 672–693.