Navigating Negativity in Research: Methodological and Ethical Considerations in the Study of Antisocial, Subversive and Toxic Online Communities and Behaviours

Kimberley R. Allison
Department of Psychology, Macquarie University
kimberley.allison@hdr.mq.edu.au

Abstract

Amidst increasing recognition of the prevalence and impact of online aggression, hate and prejudice, a rapidly expanding and multidisciplinary literature has emerged exploring the many manifestations of online negativity. Despite the varied approaches employed, these studies share common methodological and ethical issues associated with the collective focus on antisocial online activities. This paper explores issues around study design, ethics and safety, and presentation of findings which are prominent in this literature, drawing from personal and previously published research experiences in order to inform a discussion of how scholars can consider and navigate the challenges inherent in this field of study.

Introduction

Despite increasing consensus amongst the general population that online aggression and toxicity are problematic, antisocial behaviours remain persistently prevalent within mediated spaces (Duggan 2017; Duggan et al. 2014) and due to the rapidly evolving affordances of online platforms, are manifesting in increasingly varied and constantly changing forms (Corcoran, McGuckin, and Prentice 2015; Marwick and Caplan 2018). Efforts to understand, address and prevent these behaviours has expanded accordingly with research increasingly examining common forms such as cyberbullying (e.g. Cassidy, Faucher, and Jackson 2013), hate speech (e.g. Chandrasekharan et al. 2017), trolling (e.g. Coles and West 2016) and flaming (e.g. Jane 2015), as well as more recent explorations of celebrity hate (e.g. Ouvrein, Vandebosch, and De Backer 2017) , toxic communities (e.g. Hine et al. 2017) and networked campaigns of harassment (e.g. Massanari and Chess 2018).

Research into antisocial online behaviours sits at the intersection of academic disciplines, with scholars from communications, computing, and psychology among those involved in these investigations. As a result, previous work on toxic and subversive behaviours and communities has utilised varied study designs, methodologies and theoretical frameworks, informed by different disciplinary traditions. For example, scholars have surveyed and interviewed those engaged in and affected by online negativity (e.g. Blackwell et al. 2018; Rimington 2019), run focus groups and experiments exploring perceptions of different types of cyber-aggression (e.g. Allison, Bussey, and Sweller 2019, 2020; Ouvrein et al. 2017), completed ethnographies of platform cultures (e.g. Diaz-Fernandez and Evans 2019; Massanari 2015), carried out computational analyses of content posted to online communities (e.g. Mittos et al. 2020), and developed techniques to detect and classify antisocial online behaviours (e.g. Chandrasekharan, Gandhi, and Mustelier 2019). The different approaches each have their strengths, limitations and challenges, making them suitable to answer different research questions about the phenomena under study. The resulting picture of online negativity that emerges when synthesising findings is thus far richer and more nuanced than could be produced from any single approach in isolation, highlighting the importance of interdisciplinary collaboration and conversation in advancing our understanding of these communities and behaviours.

However, this work can at times be confronting and fraught with methodological and ethical issues to those involved, including participants, researchers, and those engaged in the behaviours or communities under investigation. As has been noted for HCI research more generally, scholars may find organisational and regulatory guidelines (e.g. institutional review boards, platform terms of services) ill-equipped to advise them on how to navigate challenges which arise throughout the research process (Munteanu et

al. 2015; Brown et al. 2016; Waycott et al. 2017). Instead, there has been increasing emphasis on facilitating open conversation, sharing of experiences, and learning from peers to support ethical decision-making, rather than prescribing further rules and restrictions on research (Fiesler et al. 2018; Markham and Buchanan 2012; Munteanu et al. 2015; Vitak, Shilton and Ashktorab 2016; Waycott et al. 2017).

Responding to these prompts, this paper discusses some of the challenges common to research on online negativity concerning study design, ethics and safety, and communication of findings when studying online negativity, with reference to how I and other scholars have navigated these issues in previous work. Specifically, I explore issues around (1) how research foci are conceptualised; (2) development of rapport with participants; (3) negotiation of rights and risks of involvement in studies; (4) management of risks to researchers; and (5) appropriate reporting of findings. These considerations are important in encouraging more ethical, sensitive and rigorous research into toxic and subversive online communities and behaviours in the future.

Conceptualising communities and behaviours: Defining and operationalising research foci

The majority of research in this field considers individual communities, platforms, and/or behaviours in isolation. While this restricted focus is often necessary in order to explore phenomena in depth, increasing evidence suggests that toxic and subversive communities and behaviours can be interconnected and difficult to delimit or delineate. For example, when exploring toxic discussions of genetic testing on Reddit and 4chan, Mittos and colleagues (2020) found links between similar racist and fringe political communities across the different platforms, as well as large overlaps with other prejudiced communities (e.g. misogynistic men's rights groups) on Reddit. Similarly, Topinka (2017) notes explicit links between subreddits using humour as a cloak for racism and those exhibiting other forms of prejudice. Other literature has evidenced the spread of racist memes across communities and platforms (Zannettou et al. 2017), common language and patterns of discourse across misogynist online spaces (Marwick and Caplan 2018), and "raids" of YouTube videos by 4chan users (Hine et al. 2017). It is clear that **further attention to broader toxic communities whose activity spans multiple platforms** is warranted; previous work on the loose network of online groups characterised by anti-feminist and misogynistic beliefs (dubbed the manosphere; e.g. Marwick and Caplan 2018) may be useful in informing similar approaches to other communities.

Further consideration of how behaviours of interest are defined is also needed. Academic definitions of different types of cyber-aggression (e.g. cyberbullying, trolling) are not always consistent with how those involved in these interactions perceive their experiences (Allison and Bussey 2020b), and this is true of both sources (Talwar, Gomez-Garibello, and Shariff 2014) and targets (Duggan 2017). Moreover, terms describing different forms of online negativity have been inconsistently defined across studies, making it difficult to compare and synthesise findings (Patchin and Hinduja 2015). Going forward, researchers should carefully **consider how they define and operationalise the phenomena under study, ensuring this is clearly communicated to participants**. In cases where types of online negativity are defined in terms of specific behaviours (e.g. in instruments assessing involvement in or attitudes towards cyberbullying), these **measures should be regularly reviewed to ensure newly emergent forms of negativity are represented**; ideally, these should be **developed with input from those involved in or impacted by these behaviours** (Corcoran et al. 2015). The same is true for studies using automated or computational methods to detect online negativity (e.g. hate speech), whose lexicons require regular updates to incorporate newer terms and expressions (Chandrasekharan et al. 2019). In considering their analytic methods, researchers should **consider the ability of their approaches to account for contextual nuances** in the classification of observed posts and behaviours, which may better distinguish more ambiguously toxic or harmful content (e.g. Chancellor, Hu, and De Choudhury 2018).

Recruitment and rapport: Building trust and validating participants' experiences

For scholars working directly with those involved in or affected by antisocial online behaviours, the identification and recruitment of participants may be impeded by mistrust of researchers and reluctance to report stigmatised behaviours and experiences. This is true of both sources and targets of online negativity (Barratt and Maddox 2016; Jane 2015). For example, Jane (2015) notes that women who have been targeted by flaming may experience a "*tyranny of silence*" similar to those who have experienced physical or sexual harassment or assault, with shame and fear potentially deterring them from discussing their experiences. Here, additional efforts to **secure and assure participants of their safety, anonymity, and confidentiality of disclosed experiences** may help to build trust and rapport (Barratt and Maddox 2016); this in turn requires **careful planning around how data will be securely stored and de-identified** when presenting study findings, in order to minimise the possibility of further risk or harm to participants. Despite the challenges of direct recruitment, this approach may be preferable to analysing discussions already occurring in online communities, as the presence and attention of researchers may threaten the safety and usefulness of these spaces for those affected (Williams Veazey et al. 2019).

It is important that these considerations are not tokenistic gestures primarily intended to maximise participant recruitment, retention and compliance, but genuine attempts to ensure their safety and interests are prioritised (Waycott et al. 2017). Here, **participatory research design- involving those whose experiences are the subject of study-** may help to both build trust and ensure that study methods and outputs are sensitive and appropriate for participants; these approaches are increasingly common in HCI research (Vines et al. 2015). Researchers should note, however, that members of marginalised groups are disproportionately likely to be both targets of online negativity (Blackwell et al. 2018; Duggan et al. 2014) and subjects of research (Brown et al. 2016). Additional care is therefore warranted to **ensure that research has meaningful benefits for these communities**, rather than merely co-opting their experiences for academic gain (Brown et al. 2016; Vines et al. 2015). Again, partnerships with these communities and/or organisations that advocate for them (e.g. Blackwell et al. 2018; Freed et al. 2019) may help to facilitate more tangible and immediate communal benefits and knowledge translation. Where possible, these **member representatives should be adequately compensated** to minimise epistemic exploitation (Berenstein 2016).

Issues of trust may also challenge those whose research focuses on sources of online negativity. These individuals may be reluctant to report their engagement in cyber-aggression due to social desirability biases or mistrust of researchers (Barratt and Maddox 2016) or may not consider themselves as perpetrators despite engaging in behaviours deemed toxic or antisocial by scholars (Coles and West 2016). Here, **engaging gatekeepers of these communities (e.g. moderators) as advocates for research participation** may improve recruitment (Barratt and Maddox 2016; Rimington 2018)- although for toxic communities where research attention is likely to be critical, gatekeeper buy-in may be difficult to achieve, and relations with the community may be undermined should findings be made public.

Given the nature of this research, some academics have also questioned the extent to which sources' accounts of online negativity can be trusted; cyber-aggressors may not be completely conscious of their own motivations, or may not be willing to accurately represent their experiences (Jane 2015). Issues of data integrity, common to online studies across domains, may be a particular concern for researchers investigating sources of negativity- for example, a study of trolling behaviour may be at risk of being sabotaged by those it intends to recruit. Planning ahead to **incorporate means of integrity checking data** may be useful in managing dishonest and ingenuine responses, particularly for online studies. In our previous work investigating perceptions and reasoning about online negativity (Allison et al. 2019, 2020) we elected to recruit from groups considered less likely to respond non-seriously (i.e. undergraduates, MTurk masters), and additionally incorporated open-ended questions and manually checked data to further screen out ingenuine respondents. While we acknowledge that respondents' reasoning may not be representative of those directly engaged in antisocial online behaviours, **triangulation of findings** with previous work conducted on toxic and subversive communities indicated significant parallels and overlaps, suggesting generalisability to more extreme cases.

Protecting participants: Ethical design and data collection

Managing risks associated with participation

Empirical studies on toxic online behaviours and communities carry an inherent risk of distress and discomfort for those who participate in this research, especially if this involves exposure to explicit content (e.g. Newton 2019, 2020). This may be particularly pronounced for individuals who have previous targets of antisocial online behaviours, as well as those at increased risk of being targeted because of their identity (e.g. women, people of colour, LGBTQI+ people; Blackwell et al. 2018; Duggan et al. 2014). It is therefore important for researchers to **consider and implement strategies to address potential harms to participants**, balancing the need to investigate serious forms of online negativity in an ecologically valid way with the need to preserve participants' psychosocial wellbeing.

In doing so, **study designs should both inform and be informed by the support services and resources available** to support participants. For example, our interview study exploring perceptions and reasoning about different forms of online negativity (Allison and Bussey 2020b) was able to use vignettes depicting more serious incidents to elicit discussion as interviews were conducted in person, allowing the interviewer to monitor and address potential distress as it arose. By contrast, our subsequent experimental work (Allison et al. 2019, 2020) was online and international, complicating the abilities to both detect distress and refer those impacted to appropriate support services. As a result, we elected to use only milder examples of online negativity in these studies, potentially impacting findings; exploring more serious forms may necessitate **additional reassurance around participants' rights to discontinue or withdraw** from the study without consequence, as well as **precautions for detecting and responding to potential harms as they arise** (e.g. incorporating distress screening into the study design and referring participants who score above a clinically significant cut-off score to support services).

In addition, research into the impacts of online negativity may uncover or trigger more chronic psychosocial difficulties which cannot be appropriately addressed by conventional crisis services (e.g. phone helplines). This risk may be

heightened for studies investigating impacts of online negativity on targets, particularly those who have been targeted repeatedly or by more severe acts of cyber-aggression. In these cases, it is prudent for researchers (particularly those working more closely with participants) to **be aware of means by which participants can access longer-term psychosocial support**. An exemplar of this can be seen in Freed and colleagues' (2019) study on technological abuse in intimate partner violence, which was embedded in a family justice service providing support to survivors. For researchers not affiliated with such services, the ability to advise participants on affordable and accessible support options may be an acceptable substitute. For example, when preparing to conduct interviews for our initial qualitative studies (Allison and Bussey 2020a, 2020b), the interviewer familiarised themself with a national initiative by which participants could access longer-term, subsidised mental health care.

Researchers should also **carefully consider the necessity of exposing participants to more serious or extreme forms of negativity** (Newtown 2019), avoiding unnecessary duplication of previous work and ensuring studies are designed and conducted as rigorously as possible to maximise learnings from this research. Certainly, there are cases where scholars may consider the potential benefits of their work to justify the risks associated with participation. For example, training datasets are required to build AI classifiers that may be able to automatically detect and remove offensive content from a platform, and this necessitates initial manual work to label harmful content. In such cases, there may still be ways to minimise unnecessary exposure and its associated risks- for example, if it is possible to reuse previously labelled datasets (e.g. Chandrasekharan et al. 2019) or concurrently trial means of reducing the impacts of exposure (e.g. by displaying content in black-and-white or without sound; Newton 2020). In other cases, exposing participants to extreme forms of negativity may be judged unnecessary. In our investigations of individuals' reasoning about the acceptability of potentially negative online behaviours, for example, we hypothesised that severe/graphic incidents would elicit the same kind of absolute moral judgements seen for behaviours associated with risks of serious harm (e.g. doxxing), but ultimately decided that testing this hypothesis was not worth the additional risk of exposure to this content (Allison and Bussey, 2020b).

Consent, transparency and the use of public data

Consideration must also be given to those who have not traditionally been viewed as active participants, but whose data is nonetheless used in research (Fiesler et al. 2016)- particularly for ethnographies or content analyses of online communities. The compilation of large social media datasets (e.g. through APIs or scraping) allows for high-powered analyses which are relatively unaffected by self-selection biases seen in direct research; however, the ethics and legality of such practices are contentious (Fiesler, Beard, and Keegan 2020). While many professional psychological bodies and ethics committees do not require consent for studies of behaviours in public spaces (Ethics Committee of the British Psychological Society 2009) including online spaces (Coles and West 2016; Fiesler et al. 2016), users' public posting does not necessarily mean they would expect or consent for their content to be used in research- particularly if this has potential negative consequences for the posters or their communities (Hu 2019).

Amongst HCI researchers working with public data, transparency with participants is identified as an important part of ethical academic practice- particularly around the intent and methods of data collection (Vitak et al. 2016). At the same time, notifying and obtaining active informed consent from all users whose data is used may be unfeasible, given the scale of these datasets and the transience of community membership. In considering these factors, researchers may find official ethical and legal guidelines to be of limited use- for example, ethics committees may have inconsistent policies around social media data collection and what constitutes human subjects research (Vitak et al. 2017), while platform terms of service are frequently broad and ambiguous, offering little concrete guidance around the permissibility of research (Fiesler et al. 2020). Likewise, platform users' presumed awareness and permissiveness (or lack thereof) of the use of their data for research cannot be the sole standard of ethical practice (Fiesler et al. 2020).

Particularly in the case of critical research on toxic and antisocial communities and behaviours, transparency around research motivations, questions and outputs (or in the case of analysis of public posts, its occurrence) may be undesirable due to the potential to undermine participant trust, hamper recruitment, and result in backlash that may compromise research integrity or researchers' safety. For example, those engaged in toxic online communities or targeted campaigns of harassment actively seek to deter critical studies exposing or addressing their actions (e.g. Massanari and Chess 2016). While deception, lack of informed consent, and data collection which is likely to violate community wishes or platform goals may be deemed unethical according to deontological approaches, others may consider engaging in this work to be justified in the service of countering harmful expressions of hate and aggression. For example, Fiesler and colleagues recommend that **both harms and benefits- to platforms, their users, and broader society- all be considered when evaluating the ethicality of potential methods**; this may in turn be influenced by the identities of researchers, the nature and topic of the research, and users' expectations and awareness of research (Fiesler et al. 2020; Fiesler and Proferes 2018).

Previous discussions around the ethics of presenting public posts in research (e.g. quoting Reddit comments in

published articles) has tended to focus on content relating to sensitive and stigmatised experiences and identities, such as physical and mental health issues (Haimson, Andalibi, and Pater 2016). In such cases, thorough de-identification of data is recommended, particularly if presenting this content outside of its intended context (often in more permanent and visible forms) may have potential consequences for the individuals or their communities (Fiesler et al. 2016; Haimson et al. 2016). While certainly an important consideration in these instances, this raises an interesting question in the case of research on toxic online communities and behaviours: does the anonymisation of qualitative data protect participants from harm or shield them from the consequences of their actions? Researchers seeking to protect the privacy of posters must also consider implications of the searchability of public posts (Fiesler et al. 2016); **anonymising quotes may require not only the removal of identifying (e.g. account) information, but also the rephrasing of content** so that it cannot be traced back to the original post.

Similar considerations arose in relation to our previous work: while public posts were not used as a data source, they were used as experimental stimuli to explore participants' judgement of different forms of online negativity (Allison and Bussey 2020b; Allison et al. 2019). This choice was made to improve the ecological validity of the experimental design, ensuring that stimuli reflected actual examples of the phenomena under study. Our evaluation of the risks of using these materials suggested that the likelihood of harm to those whose posts were used were minimal, as these were only viewed by participants under controlled circumstances: either in person under supervision, ensuring they would not be able to further disseminate identifiable content (Allison and Bussey 2020b), or with the more identifiable details of the posts removed (Allison et al. 2019).

To limit risks associated with further spreading this content in contexts not intended by the posters, we chose not to publish these posts in more permanent forms. Instead, when sharing study materials (e.g. in paper appendices) we have provided a summary of the essential details of these posts in a way that makes it more difficult to identify those involved, providing a greater measure of privacy without compromising methodological transparency. We would additionally recommend **seeking consent from the original posters prior to the use of their content**, although we acknowledge that this may not always be feasible (Allison and Bussey 2020b). By contrast, our other work has used artificially created vignettes (Allison et al. 2020), ensuring greater experimental control and avoiding these ethical issues albeit at the possible expense of ecological validity; these vignettes were therefore included verbatim in paper appendices, although specific (celebrity) target names were removed to limit possible harm.

Staying safe: Researchers' safety and wellbeing

Although ethical considerations in research have typically focused on potential harm to participants (and to a lesser degree, the communities they represent), increasing attention is being paid to the potential impact of this work on researchers themselves. This is reflected in both peer-reviewed (e.g. Massanari 2018) and grey (e.g. Marwick, Blackwell, and Lo 2016) literature, as well as in discussions during conference workshops (e.g. Fiesler et al. 2019) and presentations (e.g. Clancy 2019).

Of particular concern is the safety and wellbeing of researchers working with potentially hostile individuals and communities: in addition to general emotional stressors associated with potentially distressing research topics (Fiesler et al. 2019; Mitchell and Irvine 2008) which may be heightened for member-researchers (Nelson 2020) and those studying explicitly aggressive and prejudiced behaviours (Diaz-Fernandez and Evans 2019; Trott 2019), scholars researching toxic technocultures may find themselves under attack by the communities they are studying (Chess and Shaw 2015; Mariconti et al. 2019; Massanari 2018). For example, writing about critical games scholarship and research into the Gamergate movement, both Chess and Shaw (2015) and Massanari (2018) report that the discovery of critical feminist scholarship by Gamergate supporters resulted in public online attacks on the works and their authors, including targeted harassment and doxxing. This has been noted to intersect with other biases evident within the Gamergate movement, including misogyny, racism, homophobia and transphobia, and negative views towards critical social science research (Massanari 2018; Rimington 2018). Moreover, these risks are not limited to a single community or movement: Massanari (2018) notes that the far-right targets not only those who research their group, but those who threaten their values more broadly (including any issue concerning social justice, intellectualism, progressivism or liberalism).

These concerns are non-trivial: both the psychosocial impact of researching toxic communities and the potential for retaliation factored into our decision to focus on less extreme examples of negativity in the general population in our work (Allison and Bussey 2020b; Allison et al. 2019, 2020). However, the deterrence and effective silencing of dissent and critical study is argued to be precisely the objective of those engaged in targeted harassment of academics (Massanari 2018). As such, it is crucial to both legitimise and encourage the study of these communities and behaviours, while ensuring researchers are adequately supported in managing risks and impacts of this work. Suggested support strategies mirror those recommended to researchers in potentially hostile offline spaces: primarily, **encouraging careful identity management** (Rimington 2018) and stressing the importance of **professional therapeutic and peer**

support (Massanari 2018). For example, in their ethnography of toxicity in gaming communities Rimington (2018) made the decision to obscure personal information (including their gender) for safety, although they note that this presented challenges when exploring emerging themes around gendered toxicity. To a degree, this was less of an issue in our research: while I hold multiple identities that are often targeted by toxic online communities and cyber-aggressors more broadly, these are not visible to either face-to-face interviewees or online participants. However, this meant investigating perceptions of more ambiguous expressions of prejudice (e.g. the ostensibly civil, coded homophobia of arguments against marriage equality reported in our previous work) was particularly confronting and therefore avoided.

Massanari (2018) further stresses the importance of **developing research networks around toxic cultures** like the "alt-right" in order to establish the importance of this work, share learnings across communities, and buffer impacts of targeted harassment (particularly for marginalised and vulnerable groups). Indeed, this emerged as key driver of conversations at conferences about difficult research experiences, which highlighted the need for increased attention and collaboration to assess and ameliorate these impacts (Fiesler et al. 2019; Trott 2019). Institutional and systemic change is also necessary: particularly where research institutions expect public outreach and media engagement as part of research dissemination, **support systems must be implemented and ready to help researchers manage potential negative attention** arising from the increased visibility of their work (Clancy 2019; Marwick et al. 2016). **Systemic support and accommodation to allow flexibility around timelines and improve access to mental health care** may also help to support sustainable research practices and prevent burnout. This might include increasing institutional funding for counselling services, facilitating extensions of candidature for student researchers, or allowing flexibility in revise and resubmit deadlines for journal publications (although it is less clear how this would work with the rigid deadlines associated with HCI conferences). More radical **rethinking of how sensitive research is disseminated** may also be necessary- for example, reconsidering traditional sharing of findings with the communities under study, allowing anonymous or pseudonymous authorship, and potentially reworking tenure and promotion systems to accommodate these changes (Massanari 2018).

Talking toxicity: Reporting on antisocial online communities and behaviours

Finally, careful consideration is needed around how research on toxic and subversive online communities is discussed and disseminated. While scientific communication is an important skill across research areas, particular care is needed in this field given the deliberate attempts of these communities to reframe discourse to legitimise their actions (e.g. Marwick and Caplan 2018; Massanari 2018) and the potential for these discussions to amplify, publicise and normalise both specific communities and negative behaviours (e.g. Phillips and Milner 2017). In addition to potential issues around posters' consent and privacy raised previously (Haimson et al. 2016), the act of using or reproducing negative behaviours (particularly more explicit, severe or prejudiced forms) in research is ethically contentious. Verbatim replication may be interpreted as a form of symbolic violence (Massanari 2018) and may distress both participants and research audiences (Allison and Bussey 2020b; Newton 2019), while the tendency to refer to cyber-aggression euphemistically or reproduce only milder forms may misrepresent acts as more benign than their reality (Jane 2015). A potential compromise may be to replicate verbatim examples (or descriptions thereof, in cases where publishing images is illegal) but to **include content warnings when presenting work** so that audiences may prepare themselves, allowing for management of potential harm (Lockhart 2016).

While the increasing prevalence and impact of these behaviours is difficult (and potentially unethical) to ignore, even critical research and media coverage risks amplifying the visibility and influence of antisocial communities, potentially normalising and spreading these behaviours, and playing into the hands of those seeking to manipulate the discourse around these issues (Phillips 2018; Phillips and Milner 2017). Here, Phillips' (2018) **guidelines on journalistic reporting around online extremism, antagonism and manipulation may be of use to researchers** (indeed, these echo suggestions made by Massanari (2018)). These include centring perspectives of those impacted, rather than allowing sources to control the narratives; describing incidents as accurately as possible without resorting to euphemism or minimising behaviours; and exercising caution when reproducing examples of negativity (Phillips 2018).

Conclusions

Despite the varied approaches and foci of studies on antisocial online communities and behaviours, common challenges exist around operationalising constructs, working with participants and their data, and presenting findings. In each of these cases, no single "correct" solution exists as to how this research should be conducted; instead, the most ethical, responsible and sensitive solution will vary depending on the academics, research questions, and methods in each scenario. As Fiesler and colleagues (2020) note, the lack of clear, consistent or universal guidance may be frustrating for researchers facing these challenges, and this is true both of ethical and regulatory bodies and of this paper. However, such decisions are by nature highly contextually

dependent, with multiple defensible positions. Rather than prescribing a specific map for others to navigate these issues, it is hoped that this paper will encourage the careful consideration of methodological and ethical challenges in the design, execution and presentation of research; highlight the importance of establishing strong research networks to support sensitive, ethical and sustainable practices in the study of online aggression and hate; and promote more widespread and open discussion of these matters within and beyond the academic community.

Acknowledgements

Eternal thanks to Kay Bussey and Naomi Sweller for their support, mentorship, and involvement with the research projects discussed in this paper.

References

Allison, K. R. and Bussey, K. (2020a). Communal quirks and circlejerks: A taxonomy of processes contributing to insularity in online communities. *Proceedings of the Fourteenth International AAAI Conference and Web and Social Media.* Palo Alto, CA: AAAI.

Allison, K. R. and Bussey, K. (2020b). *"You can't get away with this": Lay perceptions and judgements about the acceptability of online negativity*. Manuscript submitted for publication.

Allison, K. R., Bussey, K. and Sweller, N. (2019). "I'm going to hell for laughing at this": Norms, humour, and the neutralisation of aggression in online aggression. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW): article 152.

Allison, K. R., Bussey, K. and Sweller, N. (2020). Fair game: The effects of attack subject, target identity and role-relevance in the judgement of online aggression. *Proceedings of the 11th International Conference on Social Media and Society*. New York, NY: ACM.

Barratt, M. J. and Maddox, A. 2016. Active engagement with stigmatized communities through digital ethnography. *Qualitative Research* 16(6): 701-719.

Blackwell, L. Diamond, J., Schoenebeck, S. and Lampe, C. 2018. Classification and its consequences for online harassment: Design insights from HeartMob. *Proceedings of the ACM on Human-Computer Interaction* 1(CSCW): article 24.

Brown, B., Weilenmann, A., McMillan, D. and Lampinen, A. 2016. Five provocations for ethical HCI research. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*: 852-863. New York, NY: ACM.

Cassidy, W., Faucher, C. and Jackson, M. 2013. Cyberbullying among youth: A comprehensive review of current international research and its implications and applications to policy and practice. *School Psychology International* 34(6): 575-612.

Chancellor, S., Hu, A. and De Choudhury, M. 2018. Norms matter: Contrasting social support around behavior change in online weight loss communities. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, paper 666. New York, NY: ACM.

Chandrasekharan, E., Gandhi, C. and Mustelier, M. W. 2019. Crossmod: A cross-community learning-based system to assist Reddit moderators. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW): article 174.

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J. and Gilbert, E. 2017. You can't stay here: The efficacy of Reddit's 2015 ban examined through hate speech. *Proceedings of the ACM on Human-Computer Interaction* 1(CSCW): article 31.

Chess, S. and Shaw, A. 2015. A conspiracy of fishes, or, how we learned to stop worrying about #GamerGate and embrace hegemonic masculinity. *Journal of Broadcasting and Electronic Media* 59(1): 208-220.

Clancy, L. 2019. "She is fake news": Doing impact work while female. Paper presented at Digital Intimacies 5: Structures, cultures, power. Melbourne, Australia, December 9-11.

Coles, B. A. and West, M. 2016. Trolling the trolls: Online forum users' constructions of the nature and properties of trolling. *Computers in Human Behavior* 60: 2330244.

Corcoran, L., McGuckin, C. and Prentice, G. 2016. Cyberbullying or cyber aggression?: A review of existing definitions of cyber-based peer-to-peer aggression. *Societies* 5: 245-255.

Diaz-Fernandez, S. and Evans, A. 2019. "Fuck off to the tampon bible": Misrecognition and research intimacy in an online mapping of "lad culture". *Qualitative Inquiry* 25(3): 237-247.

Duggan, M. 2017. *Online Harassment 2017*. Washington, DC: Pew Research Center.

Duggan, M., Rainie, L., Smith, A., Funk, C., Lenhart, A. and Madden, M. 2014. *Online Harassment*. Washington, DC: Pew Research Center.

Ethics Committee of the British Psychological Society. 2009. *Code of ethics and conduct: Guidance published by the ethics committee of the British Psychological Society.* Leicester, UK: British Psychological Society.

Fiesler, C., Beard, N. and Keegan, B. C. 2020. No robots, spiders or scrapers: Legal and ethical regulation of data collection methods in social media terms of service. *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media.* Palo Alto, CA: AAAI.

Fiesler, C., Brubaker, J. R., Forte, A., Guha, S., McDonald, N. and Muller, M. 2019. Qualitative methods for CSCW: Challenges and opportunities. *Conference Companion Publication of the 2019 Conference on Computer-Supported Cooperative Work and Social Computing*, 455-460.

Fiesler, C., Hancock, J., Bruckman, A., Muller, M., Munteanu, C. and Densmore, M. 2018. Research ethics for HCI: A roundtable discussion. *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems:* panel 5.

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

Fiesler, C., Wisniewski, P., Pater, J. and Andalibi, N. 2016. Exploring ethics and obligations for studying digital communities. *Proceedings of the 19th International Conference on Supporting Group Work*, 457-460.

Freed, D., Havron, S., Tseng, E., Gallardo, A., Chatterjee, R., Ristenpart, T. and Dell, N. 2019. "Is my phone hacked?": Analysing clinical computer security interventions with survivors of intimate partner violence. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW): article 202.

Haimson, O. N., Andalibi, N. and Pater, J. 2016. Ethical use of visual social media content in research publications. *Research Ethics Monthly.*

Hine, G. E., Onaolapo, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Samaras, R., … and Blackburn, J. 2017. Keks, cucks, and god emperor Trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. *Proceedings of the Eleventh International AAAI Conference on Web and Social Media,* 92-101. Palo Alto, CA: AAAI.

Hu, J. C. 2019. Should researchers be allowed to use YouTube videos and Tweets? *Slate.*

Jane, E. A. 2015. Flaming? What flaming? The pitfalls and potentials of researching online hostility. *Ethics in Information Technology* 17: 65-87.

Lockhart, E. A. 2016. Why trigger warnings are beneficial, perhaps even necessary. *First Amendment Studies* 50(2): 59-69.

Mariconti, E., Suarez-Tangil, G., Blackburn, J., De Cristofaro, E., Kourtellis, N., Leontiadis, L., … and Stringhini, G. 2019. "You know what to do": Proactive detection of YouTube videos targeted by coordinated hate attacks. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW): article 207.

Markham, A. and Buchanan, E. 2012. *Ethical decision-making and internet research: Recommendations from the AoIR ethics working committee (v2.0)*. https://aoir.org/reports/ethics2.pdf

Marwick, A. E. and Caplan, R. 2018. Drinking male tears: Language, the manosphere, and networked harassment. *Feminist Media Studies* 18(4): 543-559.

Marwick, A., Blackwell, L. and Lo. 2016. *Best practices for conducting risky research and protecting yourself from online harassment (Data & Society guide)*. New York, NY: Data & Society Research Institute.

Massanari, A. L. 2015. *Participatory culture, community and play; Learning from Reddit.* New York, NY: Peter Lang Publishing.

Massanari, A. L. 2018. Rethinking research ethics, power, and the risk of visibility in the era of the "alt-right" gaze. *Social Media + Society* 4(2): 1-9.

Massanari, A. L. and Chess, S. 2018. Attack of the 50-foot social justice warrior: The discursive construction of SJW memes as the monstrous feminine. *Feminist Media Studies* 18(4), 525-542.

Mitchell, K. J. and Irvine, A. 2008. I'm okay, you're okay?: Reflections on the well-being and ethical requirements of researchers and research participants in conducting qualitative fieldwork interviews. *International Journal of Qualitative Methods* 7(4): 31-44.

Mittos, A., Zannettou, S., Blackburn, J. and De Cristofaro. 2020. "And we will fight for our race!" A measurement study of genetic testing conversations on Reddit and 4chan. *Proceedings of the Fourteenth International AAAI Conference and Web and Social Media.* Palo Alto, CA: AAAI.

Mondal, M., Araújo Silva, L. and Benevenuto, F. 2017. A measurement study of hate speech in social media. *Proceedings of the 28th Conference on Hypertext and Social Media,* 85-94. New York, NY: ACM.

Munteanu, C., Molyneaux, H., Moncur, W., Romero, M., O'Donnell, S. and Vines, J. 2015. Situational ethics: Re-thinking approaches to formal ethics requirements for human-computer interaction. *Proceedings of the 2015 CHI Conference on Human Factors in Computing Systems*: 105-114. New York, NY: ACM.

Nelson, R. 2020. Questioning identities/shifting identities: The impact of researching sex and gender on a researcher's LGBT+ identity. *Qualitative Research.* Advance online publication.

Newton, C. 2019. The trauma floor: The secret lives of Facebook moderators in America. *The Verge.*

Newton, C. 2020. Half of all Facebook moderators may develop mental health issues. *The Verge.*

Ouvrein, G., Vandebosch, H. and De Backer, C .J. S. 2017. Celebrity critiquing: Hot or not? Teen girls' attitudes on and responses to the practice of negative celebrity critiquing. *Celebrity Studies* 8(3), 461-476.

Patchin, J. W. and Hinduja, S. 2015. Measuring cyberbullying: Implications for research. *Aggression and Violent Behavior* 23: 69-74.

Phillips, W. 2018. *The oxygen of amplification: Better practices for reporting on extremists, antagonists, and manipulators.* New York, NY: Data & Society Research Institute.

Phillips, W. and Milner, R. 2017. *The ambivalent internet: Mischief, oddity and antagonism online.* Cambridge, UK: Polity.

Rimington, E M. 2018. *The social function of toxic behavior in an online video game*. PhD dissertation. Faculty of Social, Human and Mathematical Sciences, University of Southampton, UK.

Talwar, V., Gomez-Garibello, C. and Shariff, S. 2014. Adolescents' moral evaluations and ratings of cyberbullying: The effect of veracity and intentionality behind the effect. *Computers in Human Behavior* 36: 122-128.

Topinka, R. J. 2017. Politically incorrect participatory media: Racist nationalism on r/ImGoingToHellForThis. *New Media and Society* 20(5): 2050-2069.

Trott, V. 2019. "Gillette: The best a beta can get": The resistance and propagation of hegemonic masculinity in response to Gillette's marketing campaign on YouTube. Paper presented at Digital Intimacies 5: Structures, cultures, power. Melbourne, Australia, December 9-11.

Vitak, J., Proferes, N., Shilton, K. and Ashktorab, Z. 2017. Ethics regulation in social computing research: Examining the role of institutional review boards. *Journal of Empirical Research on Human Research Ethics* 12(5): 372-382.

Vitak, J., Shilton, K. and Ashktorab, Z. 2016. Beyond the Belmont principles: Ethical challenges, practices, and beliefs in the online data research community. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work and Social Computing*: 941-953.

Vines, J., Clarke, R., Light, A. and Wright, P. 2015. The beginnings, middles and endings of participatory research in HCI: An introduction to the special issue on 'perspectives on participation'. *International Journal of Human-Computer Studies* 74, 77-80.

Waycott, J., Munteanu, C., Davis, H., Thieme, A., Branham, S., Moncur, W., McNaney, R. and Vines, J. 2017. Ethical encounters in HCI: Implications for research in sensitive settings. *Extended Abstracts of the 2017 CHI Conference on Human Factors in Computing Systems*: 518-525.

Williams Veazey, L., Johnson, A., Archer, C. and van der Nagel, E. C. 2019. Closed groups. Panel at Digital Intimacies 5: Structures, cultures, power. Melbourne, Australia, December 9-11.

Zannettou, S., Caulfield, T., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Sirivianos, M., … and Blackburn, J. 2017. The web centipede: Understanding how webcommunities influence each other through the lens of mainstream and alternative news sources. *Proceedings of the 2017 Internet Measurement Conference*, 405-417. New York, NY: ACM.