Hostility online: Flaming, trolling, and the public debate

Author: Ida Vikøren Andersen, Department of Foreign Languages, University of Bergen.
Email: Ida.Andersen@uib.no

Abstract

Whereas the amount of hostility found online increases, the scholarly interest in online hostility is decreasing. In this article, I discuss three questions central to the study of online hostility, namely 1) what role the text, the speaker's intention and the targets' perception should play in definitions of hostility, 2) whether hostility is always destructive or if it can also be productive in the public debate, and 3) how to distinguish between destructive and productive hostility? I demonstrate the difficulties in defining online hostility and argue that rather than aiming for definite definitions, we should acknowledge the situatedness of rhetorical practice and, consequently, that the effects and ethical implications of utterances depend on the situation. In doing so, I aim to contribute to incite renewed academic interest in flaming and trolling.

Introduction

In April 2020, the politician Kari Kjønaas Kjos from the Norwegian Progress Party announced her decision to leave politics after 15 years. One of the reasons – and certainly the one the mass media emphasised in their reporting of the event – was the hostility she experienced in social media. 'The hostility eats into my pleasure of working and wears me out. Society has gotten much tougher since I first became a full-time politician fifteen years ago', she was quoted saying (NTB, 2020).

Three months later, in July 2020, assistant professor, Alexander Sandtorv, at the University of Oslo, wrote his last tweet. In this tweet, he announced his retreat from the public debate and the closing of all his social media accounts. This was not something that the scholar wanted to do; he wished to continue to be an active voice in the public debate. All the negative attention to his person was, however, too much for him to handle.

These two events illustrate how hostility online today poses a threat to a well-functioning public sphere, where people freely can express and discuss their concerns, desires, opinions and ideas. While hostile communication has always circulated online, many accounts suggest that it has become far more *prevalent*, *toxic*, as well as *gendered* today, and that online hostility increasingly involves threats or acts of physical violence in offline domains (Jane, 2015). Worldwide, politicians, journalists, researchers, activists, as well as 'average' citizens, who raise their voice in public, are being targeted and harassed online. This may lead to self-censorship and withdrawal from the public debate. As such, online hostility threatens individuals' abilities to exert their freedom of speech and their possibilities to participate in the public debate. That is not only a problem for the individual; it is a problem for our democracy.

Expounding beyond the Norwegian context to consider online rhetoric globally, the article discusses how to identify, define and evaluate online hostility – and the particular consequences online hostility has for the public debate. I review central parts of the literature on online hostility and engage examples of online rhetoric to discuss: 1) what role the text, the speaker's intention and the targets' perception should play in definitions of hostility, 2) whether hostility is always destructive or if it can also be productive in the public debate, and – if so – then 3) how to distinguish between destructive and productive hostility? These three questions are all central in the literature on online hostility. They reveal fundamental challenges to the study of online communications, not just in relation to hostility, but in general. Furthermore, these are questions that are central to how we, from a rhetorical perspective, understand human communication and communicative processes.

While many different types of online behaviours can be hostile – some of them illegal according to law, others not – I will focus especially on two concepts that have been central in the literature on online hostility, as well as in the public debate, namely *flaming* and *trolling*.

I demonstrate the difficulties in defining online hostility and argue that rather than putting our efforts into developing definite definitions, allowing us to immediately recognise hostile utterances, we need to acknowledge the situatedness of rhetorical practice and, consequently, that the effects and ethical implications of utterances depend on the situation. The article, then, contributes to the already vast scholarly literature on online hostility that, despite an increase of online hostility, has decreased in recent years (Jane, 2015). The article, moreover, aims to contribute with new perspectives to the public conversation, in which online hostility is an increasing concern for governments, media professionals, as well as 'average' citizens.

Defining online hostility, flaming and trolling

Here, I use the concept of 'flaming' as an overall category for hostile online *communication.* Whereas the broader category of 'online hostility' encompasses both such hostile *communications* and other acts of hostility, such as hacking, catfishing, sharing and downloading illegal pornography, I reserve the term 'flaming' for *discursive* hostility. Moreover, some have argued that flaming should be seen not exclusively as an *online* phenomenon, but rather as a 'communicative episode fundamentally independent of […] the communication channel' (O'Sullivan & Flanagin, 2003, p. 76). When I, here, use the concept of 'flaming', I refer to hostility taking place in online environments and 'hostility' to refer to hostility in general – online, as well as offline. Here, then, I understand 'flaming' as *hostile online communication.*

This, relatively, broad discursive category, features several different sub-categories, among them trolling (others being, for instance, cyber-bullying and hate speech), which I argue should be understood as the online posting of hostile content, to bring about an aggressive reaction from the other. This is done primarily for the fun of it, or as the trolls themselves commonly call it: for the 'lulz', which should be taken to mean 'laughter at someone else's expense' (Phillips, 2015). Shortly, I will problematise such definitions of rhetorical practices, i.e., definitions based in the *intention* of the speaker, but for now let us understand trolling as a particular form of flaming, namely, as a form of hostile communication, motivated by a wish to provoke a reaction – a reaction by which the victim of trolling makes himself the butt of the troll's joke.

Whereas the concept of 'trolling' seems to be increasingly used – and studied, the concept of 'flaming' seems to be on the retreat in scholarly work. The majority of research on 'flaming' is quite old, considering the rapidly changing nature of online environments. Most of this work is conducted prior to the advent of Web 2.0, and before the emergence of social network sites. According to Jane (2015), flaming is largely overlooked and ignored in research on online communication today. She points at the absence of the topics of flaming and trolling in Springer's *The International Handbook of Internet Research* (Hunsinger et al., 2010) and *The Oxford Handbook of Internet Studies* (Dutton, 2013), as well as only passing mentions of the topics in Wiley-Blackwell's *The Handbook of Internet Studies* (Consalvo & Ess, 2011). While relatively absent from the handbooks of internet studies, however, the concept of trolling is increasingly gaining traction among scholars, who have examined trolling in several different online environments, including feminist discussion forums (e.g., Herring et al., 2002), gaming forums (e.g., Cook et al., 2018; Hilvert-Bruce & Neill, 2020; Thacker & Griffiths, 2012), Wikipedia (Shachaf & Hara, 2010), 4chan and Facebook (e.g., Phillips, 2011, 2015).

The academic interest in 'flaming' has – after a period of high interest in the late 80s and 90s - however, declined. Here, I wish to draw attention to three central debates in this early academic literature on flaming that I argue can assist us in understanding why that is. The bulk of this literature is located in the field of social psychology, and revolves around three debates, concerning: 1) how to define and conceptualise 'flaming'; 2) whether 'flaming' is the result of digital communication technologies or social contexts, and 3) whether flaming is irrational, flawed, and troublesome or rational, effective and productive human behaviour.

Conceptualising flames

The absence of conceptual agreement, I suggest, is a central explanatory factor for the decline in academic interest in flaming. Broadly speaking, flaming has been defined in three ways: It has been defined by reference to features of the *texts;* it has been defined by reference to the *speaker's intention,* and it has been defined by reference to the *target* and/or, other *audiences' perception.* In the literature on flaming, a shift is observable over time: Whereas early accounts of flaming commonly defined flaming based in the textual elements, later definitions have commonly emphasised the speaker's intention and the target's experience as the central defining element of a flame. What the definitions share, is that flaming is equated to disinhibited (online) behaviour. Apart from that, the definitions are commonly vague, broad and ambiguous.

Definitions located at the *textual* level, have conceptualized 'flaming' as everything from curse words and 'typographic energy' (i.e., the use of exclamation marks and capital letters), to vague notions of criticism, emotional expressions and norm violation. Some definitions can illustrate the conceptual confusion: Flaming has been defined as discourse that includes negative affect, profanity and 'typographic energy' (Lea et al., 1992), 'antisocial interaction' (Thompsen, 1996), 'an insult' (Herring, 1996), 'emotional outbursts' (Korenman

& Wyatt, 1996), 'hostile verbal behavior' (Thompsen & Foulger, 1996), 'verbal aggression', 'blunt disclosure', and non-conforming behavior' (Parks & Floyd, 1996).

Advocating for intention- and perception-based definition of flames, many scholars – many of them also located within the field of social psychology, but also within communication and media studies – have problematised definitions that are based in the features of the text alone. With good reason. Not only are many of the definitions vague and ambiguous, they are also, I argue, treating many vernacular rhetorical forms that are not necessarily hostile, as hostile – and vice versa. Let me illustrate this by offering some examples that could typically be found in comment sections online (for a discussion of the rhetorical strategies characteristic of social media, see Andersen, 2020):

> 1) 'I'm so fucking tired of all the numbskulls who comment without reading the article! At least try to substantiate your infantile claims!!'

> 2) 'You're not Norwegian, are you? I hope you get sent out of the country.'

In many early studies of flaming, where researchers commonly set out to code instances of flaming in various discussion forums, to examine how widespread flaming was in different online environments and whether flaming was more widespread within some social contexts than others (e.g., Dubrovsky et al., 1991; Kayany, 1998; Kiesler et al., 1984; McCormick & McCormick, 1992), the first of these two utterances would probably be classified as a flame. And perhaps the second one would not (see also, O'Sullivan & Flanagin, 2003, pp. 72-74).

But is the first utterance *really* hostile? And is the second one not? The first contains swearing and invectives. It calls other people 'numbskulls'. It is written in a hostile tone. But is it hostile toward another *person?* Or is it rather sanctioning the utterances of others –

utterances that, according to the speaker, are not adhering to good debate standards? The second utterance does not contain swearing, emotional outbursts or typographic energy. It consists of a (rhetorical) question: ‘You’re not Norwegian, are you?’ and expresses the speaker’s personal hope for the future: ‘I hope you get sent away’. We might, however, assume that this is both intended and experienced as a hostile personal attack targeting the nationality or ethnicity of the other. As such, it a more hostile utterance than the first, as it targets the *person,* not the issue or the person’s *utterances.*

This demonstrates the difficulties in determining whether an utterance is hostile or not, based on a set of criteria for what constitutes hostile elements in a text. It is not surprising, then, that some scholars have turned to other ways of conceptualising flames, namely by reference to the speaker’s intention and/or the target’s perception of hostility.

One especially influential account, that advocates an approach to both the speaker’s intention and the audience’s perception, in determining whether something is or is not a flame, is O’Sullivan and Flanagin’s article: *Reconceptualizing ‘flaming’ and other problematic messages* (2003). They define a ‘flame’ as ‘a message in which the creator/sender intentionally violates interactional norms and is perceived as violating those norms by the receiver as well as by third-party observers’ (p. 85). Moreover, as the title of their article suggests, they distinguish between ‘true’ flames and ‘other problematic messages’. While their definition of flames emphasises both the speaker’s intention, the target’s perception and third-party observer’s perceptions of the message, their definition of a ‘true’ flame, prioritises the speaker’s intention. ‘True’ flames, they argue, are, ‘intentional (whether successful or unsuccessful) negative violations of (negotiated, evolving, and situated) interactional norms’ (p. 84).

Studies have suggested that – although some flamers *do* intentionally offend and hurt others – much of what might be experienced as flaming by others, is rather an immediate, affective reaction (Cheng et al., 2017) or is meant or functions to express disagreement, affect or is a form of humorous play (Moor et al., 2010). As such, it does make sense to privilege the speaker's intentions, over the audience's experiences.

We could, however, turn this around and privilege the target's experiences of the utterance. An utterance may be experienced as hostile also when the speaker's intention was not to harm. Moreover, if an utterance is *experienced* as harmful it *is,* necessarily, harmful – it has caused someone harm. Jane (2015) has, therefore, suggested that the experience of flame targets should be privileged over the experience of flame producers when determining the function of and evaluating the ethical implications of online discourse.

Evaluating utterances based on how these are perceived is, however, not straightforward. A recent study found that 7.2 per cent of a representative sample of the Norwegian population had received what they perceived as hate speech online (Fladmoe & Nadim, 2017). Most of the reported utterances were, however, not directed towards protected grounds (i.e., ethnicity, nationality, skin colour, religion, disability or sexual orientation) and are, thus, not hate speech according to legal definitions (cf. Waldron, 2012). Rather, many of the utterances perceived as hate speech were directed at the content of the target's claims and arguments (Fladmoe & Nadim, 2017, pp. 58-59).

Having studied sanctions and negotiations of norms in Facebook comment section debates, I, similarly, find that most of what is sanctioned for being personal attacks in these debates is criticism directed towards the claims and arguments and, thus, what we could call 'adequate answers' (cf. Kock, 2018) in a public debate (Andersen, 2020). This suggests that the popular comprehension of what is hostile online may be so broad that too many utterances will be labelled hostile if we are to base our evaluations solely on the audience's perceptions.

An obvious problem, if we define 'flames' by reference to the speaker's intention and/or the target's perception is, then: 'When the producer's intention and the target's perception of an utterance differs – whose definition do we privilege? As rhetorical scholars have often underscored: Rhetoric *does* something to an audience, but this *something* is seldom – if ever – the same to all audiences. Utterances will always carry different meanings for different audiences – and have different effects on different audiences. Defining utterances based on the effect it has on its audience is, therefore, not straightforward. It is, however, not straightforward to define an utterance based on the speaker's intention either. Rhetorical messages are not always crafted with a clear aim of what the utterance is to do – they may be uttered as an in-the-heat of the moment, highly affective reaction. And utterances that *are* crafted with a clear intention, do not always do what they are intended to do – and they do not do the same to all people. In other words: Utterances that are not intended to harm someone might harm someone – and vice versa. An utterance that is experienced as hostile by some – may not be experienced as such by others. And besides, how can we know what someone – other than ourselves – *actually* meant and felt?

To detect and evaluate hostility online, we should rather, I argue, take the text as our starting point. We should, however, not only look at the features of the texts, rather we need to examine the texts and their context.

Having discussed all the problems related to existing definitions of flaming, you might now expect me to offer a good, clear-cut definition that can guide our interpretation of texts in context. I will not offer such a definition. As the discussion of various ways of defining hostility online has demonstrated, there are good reasons to caution against trying to reduce the complexity of online practices by offering clear, simple and universally valid definitions. These practices are always relative, interactional, negotiable and context-dependent and, therefore, I will rather advise us to understand and approach them as such. A rhetorical, humanistic perspective may, in this regard, be of great value, when examining hostility online. This perspective emphasises how utterances are situated and contingent (Kjeldsen, 2008). It stresses that the context of communication guides both the speaker's actions and the audiences' interpretations of these. Consequently, utterances – and their potential functions – should be seen as inseparable from the context in which they occur. By approaching the phenomenon of online hostility from this perspective, we acknowledge that we cannot – and should not – strive to offer definite conclusions on how an utterance functions. Rather, practices and norms are continuously being negotiated through human interaction and, consequently, we should also continuously describe and evaluate these practices as they occur in different contexts.

Digital media: Hostile rhetorical cultures?

The second debate that has been central in the literature on online hostility concerns whether 'flaming' is the result of digital communication technologies or social contexts. The first position regards negative behaviours as a consequence of the lack of social cues and restraints in digital environments, where anonymity and the absence of non-verbal cues, such as body language and tone, is thought to elicit asocial and unrestrained behaviour (Kiesler et al., 1984; Kiesler et al., 1985; Lapidot-Lefler & Barak, 2011; Suler, 2004). The latter position sees flaming as social-context dependent, i.e., as a consequence of particular topics, the participants' background and affiliation, the different forums for discussion, rather than a characteristic of the medium (Kayany, 1998; Lea et al., 1992; O'Sullivan & Flanagin, 2003; Spears & Lea, 1992).

Flaming, I argue, should be seen as a consequence both of the technology and the particular social context in which it appears. More precisely, digital technology should be seen as a central part of the social context of utterances as different digital media affords certain rhetorical actions, while obstructing others (Miller & Sheperd, 2004; Vatnøy, 2017).

At the same time, digital technology is, in contrast to at the time when much research on flaming was conducted (i.e., in the late 80s and 90s), not something fundamentally new and exterior to human behaviour. Rather, digital technology has become such a natural part of most people's everyday lives that we have become one with the technology (Hess, 2014). While still present, anonymity and the lack of non-verbal cues are, moreover, far from the only ways in which we communicate online today. Research suggests that users of social network sites often dismiss notions of anonymity and multiple identities in favour of self-expression and bodily identity (Boler, 2007; Kennedy, 2006), and today's online communication happens through written text, images, sound and video alike.

While the affordances of digital media, direct the users' practices, by making some actions more attainable than others, the actual use of the technology depends on the users' ability and willingness to utilise these affordances (Davisson & Leone, 2018, p. 88; 92). As such, the technology's affordances do not determine, but shape 'the possibility of agentic action in relation to an object' (Hutchby, 2001, 444). As such, digital technology per se does not elicit asocial and unrestrained behaviour but can, indeed, facilitate it. For instance, the logics of social media are often said to favour issues that are emotive, moral and controversial (Barberá et al., 2015; Eberholst & Hartley, 2014; Tenenboim & Cohen, 2015), and debates concerning these issues are, in turn, found to be dominated by flaming and polarisation, rather than deliberation (Janssen & Kies, 2005; Wales et al., 2010). There are, however, differences between different platforms. Each platform presents a different social context due to its unique affordances, user base, and the particular topics, emotions and genre conventions that are regarded to be appropriate by the users of the platform (O'Riordan et al., 2012; Vatnøy, 2017; Waterloo et al., 2018).

Here, I wish to change the perspective from debating whether flaming is a consequence of digital technology or social contexts, to draw attention to how online hostility may be seen as part of a particular 'internet culture', characterised by overwhelming irony and detached laughter, and where a hostile, yet playful tone, is not only tolerated but expected. In other words: A rhetorical culture in which hostility is the norm. By 'rhetorical culture', I here refer to the configuration of a typical communication situation, world view – including both a view of truth, morality, humanity and language – and a particular rhetorical style . A rhetorical culture is, thus, constituted by who speaks and who listens, cultural and technological conditions, the language available under these conditions, as well as the norms and conventions that govern the communication (for an extended discussion of the concept of 'rhetorical cultures', see Andersen, 2020, p. 4ff).

The rhetorical culture that has been called the 'internet culture', should according to Phillips (2019), be dated to the period between 2008-2012. This rhetorical culture 'aligned with and reproduced the norms of whiteness, maleness, middle-classness, and the various tech-slash-geek interests stereotypically associated with middle-class white dudes' (Phillips, 2019, p. 2). In describing the norms of this rhetorical culture, Phillips also describes the speakers' and audiences of this culture. They are predominantly young, white, technically skilled, middle-class dudes, who communicated and 'hung out' online, prior to the advent and extensiveness of social media, when suddenly all of us started hanging out on the internet. This internet culture has a rich tradition of spectacle and transgression (Coleman, 2012, p. 101), and an important norm in this internet culture is that everything is 'just for the lulz', i.e., laughter at someone else's expense (Phillips, 2011, 2015, 2019).

This resonates well with the world view of the participants in this rhetorical culture, which has been extensively discussed by Whitney Phillips (2015, 2019) and Gabriella Coleman (2012), in particular in relation to problematic online behaviours, such as trolling and hacking. Trolls and hackers' attitude to the world are summed up in the phrase: Everything is "just for the lulz', i.e., everything is just for the fun of it. Moreover, they commonly distinguish between the online world – and the online self – and the real world – and the real self. Who they are, and what they do, in the online world do not represent who they are in the real world but is fundamentally disconnected from their real, offline selves. Finally, the participants in the internet culture display an extreme, selfish understanding of liberty – in particular of freedom of speech: They understand freedom of speech as liberty for *me,* without concerns for how one's actions and utterances may impose on other's freedoms.

The increased access to the internet in the early 2000s and the emergence of many popular social network sites, made the masses join the internet. From previously being a place in which the technically savvy and particularly interested communicated, the digital sphere now became inhabited by a bunch of people who were not familiar with the rhetorical culture and its norms. Those who were familiar with the codes, then, took advantage of this so that they could laugh. In particular, the ways in which social network sites encourages self-centeredness and the authentic, heartfelt expression, is something that the trolls, according to Phillips (2015), have taken advantage of. They see an opportunity in this ego investment and emotional sensitivity: By trolling people acting this way online – they can produce laughter for themselves. As both Phillips (2015) and Coleman (2012) have pointed out, such trolling may also be a way to protest against this tendency of narcissism in our current culture and to signal that the online world is their world, not ours. The hostile behaviours of trolls could be seen as a 'virtual fence adorned with a sign bearing the following message: 'keep (the hell) out of here, this is our Homeland' (Coleman, 2012, p. 113).

What also happens – over time – is that the norms of this rhetorical subculture are cultivated and taken up in mass culture. Just as the punk and the jeans became mainstream, trolling and other forms of hostility have also been taken up in mainstream culture. Trolls are no longer a subculture of technically savvy nerds, rather – as suggested by the title of an experimental study conducted by researchers in computer science – '*anyone* can become a troll' (Cheng et al., 2017, italics added). Flaming has been described as the *lingua franca* of the internet (Jane, 2014a, 2014b; Jane, 2015), and John Hartley's (2010) term 'silly citizenship' have become influential in describing the sort of detached, ironic and playful forms of participation facilitated online.

Such descriptions point to how hostility has – by a long way – become normalised online and is not only tolerated but expected in online communications. Hostility is, in other words, a norm, developed in a marginalised rhetorical culture, that now seems to have been taken up in mainstream culture and constitutes a threat to a well-functioning public sphere. From an initial celebration of democratising potential of participatory media, which made large-scale many-to-many communication possible (Benkler, 2006; Hindman, 2009; Shirky, 2008), media organisations and journalists, global organisations, national governments and political actors, as well as scholars, now increasingly raise concerns about social network sites' and other digital media's influence on the public debate.

Destructive, harmless or productive hostility?

The article opened with some recent examples of what the consequences of the normalisation of online hostility may be, namely that scholars, politicians, journalists, activists, as well as 'average' citizens, may self-censor or entirely retreat from the public life, due to experiencing much online hostility. Other people who haven't yet, but might want to, engage in the public debate and voice their opinion publicly, might be obstructed from doing so, as they see that they, in doing so, run the risk of being harassed, hated and threatened online.

In the beginning of the article I, moreover, suggested that the third central debate in the literature on 'flaming' concerned whether this hostility is destructive or productive (Lea et al., 1992) – a dispute that may follow directly from the absence of conceptual agreement, discussed earlier. In early accounts of flaming, the phenomenon was predominantly treated as a negative and problematic consequence of new communication technology. Later, more positive –

sometimes even celebratory – accounts of flaming as either harmless or productive uses of digital technology, have dominated.

I argue that online hostility should not be seen as *either* destructive or productive, but rather as *both* destructive and productive. I advocate an approach to online hostility, that underscores the situatedness of rhetoric, and consequently argue that the effects and ethical implications of utterances depend on the situation.

I begin by outlining some potentially negative consequences of online hostility. First, as the two examples retold in the introduction illustrated, one problem is that people might refrain from participating in the public debate. Studies of the users of social media show that the fear of being subjected to hostile remarks and behaviours is an important reason for why many refrain from sharing their opinions publicly and abstain from participation in public debates online (Kruse et al., 2018; Thorson et al., 2015; Vromen et al., 2016). This is a problem for the individual, who is obstructed from using his or her freedom of speech. It is also a problem for democracy, as we run the risk of getting a public debate, where some views and perspectives are absent. Considering the direction of much hostility, which is often gendered and racially charged – more commonly performed by white men, and targeting women and minorities (Fladmoe & Nadim, 2017; Jane, 2014a, 2014c; Phillips, 2015), the consequence of people abstaining from participation in the public debate due to a fear of harassment, might be a skewed public debate, in which the voices of some groups are lesser represented than others.

A second problem that we face, is that discrimination, hatred and hostility is increasingly normalised. If hostility is the *lingua franca* online – if it is something we not only accept, but also expect – then we could contribute to the normalisation, and thereby, possibly also increase, of hostile behaviour – both online and in the worst case, also offline. In the mass media, as well as in the early academic works on flaming, there has been a tendency to take this position to online hostility. In particular, the mass media has tended to sensationalise the dangers of the internet and ascribe this with grave consequences. In particular, the view that hostile utterances may incite physical violence has commonly been voiced (Jane, 2015, pp. 73-74; Phillips, 2013).

Perhaps as a reaction to this, more recent academic works have tended to downplay the harmful consequences of online hostility, and rather examine the possibly productive outcomes of it. As for example Esther Milne has argued, in a critique of the tendency to always treat flaming as something negative, such interpretations of flaming 'do not allow any scope for exploring the *productive* or creative capacity of flaming within group and individual identity formation' (Milne, 2010, p. 172, italics in original).

A common understanding of online hostility today is that much of what might seem like hostile is misunderstood online humour, or a way of being on the internet that does not have, or is not meant to have real, in-life implications. For instance, Phillips (2015), argues that trolls view their hostile behaviour as something that is 'just for fun' online, rather than actions that are meant to cause harm in the real world. Similarly, I have argued that many instances of hostile behaviours in comment section debates on Facebook, should be understood primarily as rhetorical performances aimed at securing the final say for oneself and suggest that this is a way to safeguard one's authentic expression from the intervention of others (Andersen, 2020, p. 239ff). I have, thus, suggested that this type of hostility is without external consequences – it serves a function *within the interaction* – but is unlikely to have consequences outside of the interaction.

While some have argued that online hostility is harmless, others have argued that it is – not only harmless – but even a productive, and necessary, resource in the digital public sphere. The main arguments in these accounts can be read as echoing much rhetorical and political thinking on the role of confrontation of conflict in the public sphere, of hostility as an integral part of political argumentation, and of insults and invectives as rhetorical means for sanctioning violations of the community's norms and values (e.g., Amossy, 2010a; Jørgensen, 1998; Lund, 2012). Moreover, there is an underlying assumption here, that directness – even hostility – is a more effective and ethical mode of confrontation in the public debate, than politeness and civility. Because political rhetoric is concerned with choices between conflicting choices, values and beliefs (cf. Kock, 2009), too much civility obscures the public debate. It does so, by concealing the conflicts and the affections, with which these are held. Moreover, when I criticise you, while still being polite about it, I make myself untouchable. You cannot fight back; you cannot come up with a counterattack. I was – after all – polite. As such, hostile criticism puts adversaries on equal terms, whereas politeness obstructs the object of criticism from 'fighting back'.

Returning to the hostility we can find online, different scholars have found it to be productive both at a micro-, meso-, and macro-level of the public sphere. At the micro-level, hostility has been seen as a way for the individual to be creative, it serves entertainment purposes, it educates, and it is a resource in the formation and enactment of individual identity (Lange, 2007; Postmes et al., 2000; Vrooman, 2002; Wang, 1996).

At the meso-level of the public sphere, hostility has been said to serve as a resource in the formation of group-identity, relationship-building, as well as in upholding a community's norms and values (Douglas, 2008; Kuntsman, 2007; Postmes et al., 2000). Moreover, flame-producing trolls are said to protest against and may, thereby, contribute to renew our culture. Phillips (2013, 2015) argues that trolling may be understood as a protest against the mass media's tendency to sensationalise events, as well as a social media culture, in which authentic self-expression and narcissism prevail. Coleman (2012, p. 102), furthermore, argues that the hostile behaviours of trolls (and hackers) could be seen as a way to protect the digital sphere – the troll's and the hacker's homeland – against easy comprehension and 'cultural co-optation and capitalist commodification that so commonly prey on subcultural forms'.

At the macro-level of the public sphere, hostility can be seen as a necessary, and inherent, part of the digital public sphere and digital citizenship. Here, I will outline two main claims about the productivity of hostility in the digital public sphere. The first claim is that online hostility is a productive provocation and an integral part of agonistic digital citizenship. The claim is grounded in the assumption that conflict and confrontation are not only inevitable but necessary in a well-functioning public sphere and that, even the most hostile, verbal expressions of conflict are, in the long run, what ties us together as publics (cf. Mouffe, 2005). In other words: Even when we are fighting, mocking, and hating, we are maintaining contact with the 'other' – we are relating to one another – and this is better than ignoring or repressing one another.

This claim is promoted in Amossy's (2010b) article on flaming as a discursive and argumentative phenomenon pertaining to polemical discourse. Flaming should, according to Amossy, be understood as verbal expressions of social and political conflicts that exist beyond these online spaces and serves as a peculiar form of conflict management in society. Amossy (2010a, 2010b) views polemical discourse, and thereby also flames, as a legitimate argumentative mode and tone of voice in the public sphere. Verbal violence, in the form of polemics, she argues, is a way – in fact, it is the *only* way, for adversaries to co-exist within a shared social space. It is the only way for adversaries to 'live together while sustaining opposing claims and holding antithetical views' (Amossy, 2010a, p. 59) Verbal violence, she argues, supersedes and prevents physical fighting and political repression, as the inherent conflicts are channelled through discourse, relating people to one another, while upholding their dissent and disagreements.

Similar to this argument, is McCosker's (2014) take on online hostility as an integral part of agonistic digital citizenship. He views flames as provocations that play a legitimate and productive role in pluralistic, agonistic societies. In a similar manner to Amossy's take on the polemic, he underscores the role of contact between adversaries – even when hostile – in forging bonds of identification. He argues that hostility online is best understood in an 'agonistic sense of searching for an adversary and trying to best them, or at least maintain contact; and in response, to shore up bonds of identification and agreement' (McCosker, 2014, p. 215).

I have now outlined some conflicting positions on the place of online hostility in the public debate. I have argued that it can be destructive as it may prevent people from participating, and thus may result in a public debate where not all voices and perspectives are represented. I, then, went on to argue that much online hostility may just be 'harmless fun', or serve other purposes *within* the particular interaction, but that does not have external consequences. Finally, I suggested that online hostility may even be productive in the public sphere – both on an individual level, for the group, and as a way for people with conflicting views, beliefs and values to be able to co-exist in our shared social and cultural space. I do not privilege any of these positions over others, rather I argue that online hostility may be destructive, harmless and productive – and whether it is this or that, depends on the context of the utterance.

How, then, can we distinguish between hostile utterances that are damaging to individuals, groups and democracy – and hostile utterances that are without real, in-life consequences – or that are even productive?

How do we proceed from here?

I have grounded my argument in the understanding of utterances as situated and context-dependent*,* and will necessarily, call for approaches to online hostility that are able to take both utterances – and the context of utterances – into consideration. This is challenging for all studies of communication between human beings but is something that becomes especially challenging when we are dealing with communication in digital environments. Online both texts and contexts are complex, changing and incalculable (Kjeldsen, 2008; Hess, 2018). Communication online is characterised by intertextuality, inconsistency, immediacy, pace and fragmentation (Kjeldsen, 2008; Warnick, 1998, 2007). Contexts are countless and indefinite online, and the lines between different contexts, for instance between a private and a public context, are not clear-cut. Texts and discourses are fragmented and circulated – often far removed from their original context and speaker. Fragments from rhetorical utterances are shared and re-contextualised, commonly in multimodal utterances, such as 'memes' or 'mash-ups'. Thereby, texts gain new meaning, and in many cases, audiences only encounter the re-contextualised version of a text. Moreover, audiences are often invisible (boyd, 2008), and the communication (at least when written), is stripped of non-verbal cues that can guide of interpretation.

Moreover, genre conventions and communicative norms are constantly being challenged, contested and negotiated (Andersen, 2020, p. 271ff). What may be a norm to one participant in an interaction, might be a norm-violation for the other. This makes it extremely challenging – something that I think the literature on 'flaming' shows – to distinguish hostility from utterances that carry other functions – and to distinguish between hostility that is harmful and critique-worthy, hostility that is just 'harmless fun', and hostility that may even be productive in the public sphere. This does, however, not mean that we should not try.

The most productive way forward, I argue, is to assume perspectives on human communication found in the humanities – and to continuously describe and evaluate these practices, as they occur in different contexts. Rhetorical theory has always been concerned with how human communication and experience is situated and context dependent. As such, the rhetorical perspective has shown an ability to deal with fragmented and complex situations, and to adjust to changing technological circumstances, and changing rhetorical practices and norms.

We should continue to examine the texts and contexts*,* in which hostility occurs – or may occur – online, and we should always examine both texts and contexts together. We should not attempt to offer any definite, universal discourse ethics, that allows us to – once and for all – determine whether certain utterances are hostile or not and whether this hostility is destructive, harmless or productive. Rather, we must remember that how an utterance functions – and the ethical implications of that – always depend on the situation in which it occurs.

About the author

Ida Vikøren Andersen is a post doctor at the CLIMLIFE-project at the Department of Foreign Languages at the University of Bergen. She has a PhD from the Department of Information Science and Media Studies, University of Bergen.

Acknowledgements

The article is a revised version of my trial lecture, with the same title, given as part of the public defence of my doctoral thesis at the University of Bergen in 2020. I want to thank the two opponents, Celeste Condit and Aaron Hess, for assigning me this topic, allowing me to critically discuss online hostility.

Literature

R. Amossy, 2010a. "The Functions of Polemical Discourse in the Public Sphere. In: M. Smith & B. Warnick (editors.) *The Responsibilities of Rhetoric.* Waveland Press, pp. 52-61.

R. Amossy, 2010b. "Polemical Discourse On The Net: 'Flames' In Argumentation", *Proceedings ISSA 2010,* http://rozenbergquarterly.com/issa-proceedings-2010-polemical-discourse-on-the-net-flames-in-argumentation/.

I. V. Andersen, 2020. *Instead of the deliberative debate: How the principle of expression plays out in the news-generated Facebook discussion.* Doctoral Thesis, University of Bergen.

P. Barberá, J. T. Jost, J. Nagler, J. A. Tucher & R. Bonneau, 2015. "Tweeting From Left to Right: Is Online Political Communication More Than an Echo Chamber?", *Psychological Science, 26*(10), pp. 1531-1542.

Y. Benkler, 2006. *The Wealth of Networks. How Social Production Transforms Markets and Freedom.* Yale University Press.

M. Boler, 2007. "Hypes, hopes and actualities: new digital Cartesianism and bodies in cyberspace, *New Media & Society, 9*(1), pp. 139-168.

d. boyd, 2008. *Taken out of context: Americal teen sociality in networked publics.* Doctoral thesis, Berkely, University of California.

J. Cheng, C. Danescu-Niculescu-Mizil, J. Leskovec & M. Bernstein, 2017. "Anyone Can Become a Troll", *American Scientist, 10*(3), pp. 152-155.

G. Coleman, 2012. "Phreaks, hackers, and trolls: The politics of transgression and spectacle". In: M. Mandiberg (editor), *The social media reader,* NYU Press, pp. 99-119.

M. Consalvo & C. Ess, 2011. *The handbook of internet studies* (volume 14). John Wiley & Sons.

C. Cook, J. Schaafsma & M. Antheunis, 2018. "Under the bridge: An in-depth examination of online trolling in the gaming context, *New Media & Society, 20*(9), pp. 3323-3340.

A. Davisson & A. C. Leone, 2018. "From Coercion to Community Building: Technological Affordances as Rhetorical Forms". In: A. Hess & A. Davisson (ediotors), *Theorizing Digital Rhetoric,* Routledge, pp. 85-97.

K. Douglas, 2008. "Antisocial communication on electronic mail and the internet". In: E. A. Konijn, S. Utz, M. Tanis & S. B. Barnes (editors), *Mediated Interpersonal Communication,* Routledge, pp. 200-214.

V. J. Dubrovsky, S. Kiesler & B. N. Sethna, 1991, "The Equalization Phenomenon: Status Effects in Computer-Mediated and Face-to-Face Decision-Making Groups", *Human-Computer Interaction, 6*(2), pp. 119-146.

W. H. Dutton, 2013. *The Oxford Handbook of Internet Studies*. Oxford University Press.

M. K. Eberholst & J. M. Hartley, 2014. "Agency and civic involvement in news production via Facebook commentary". Paper presented at ECREA-conference, Lisboa, Portugal, 12-15 November.

A. Fladmoe & M. Nadim, 2017. "Silenced by hate? Hate speech as a social boundary to free speech". In: A. H. Midtbøen, K. Steen-Johnsen, & K. Thorbjørnsrud (editors.), *Boundary Struggles: Contestations of Free Spech in the Norwegian Public Sphere*, Cappelen Damm Akademisk, pp. 45-75.

J. Hartley, 2010. "Silly citizenship", *Critical Discourse Studies, 7*(4), pp. 233-248.

S. C. Herring, K. Job-Sluder, R. Scheckler & S. Barab, 2002. "Searching for safety online: Managing 'trolling' in a feminist forum", *The Information Society, 18*(5), pp. 371-384.

S. C. Herring, 1996. "Two Variants of an Electronic Message Schema". In: S. C. Herring (editor), *Computer-mediated Communication: Linguistic, Social and Cross-cultural Perspectives,* John Benjamins, pp. 81–106.

A. Hess, 2014. "You Are What You Compute (and What is Computed For You): Considerations of Digital Rhetorical Identification", *Journal of Contemporary Rhetoric, 4*.

Z. Hilvert-Bruce & J. T. Neill, 2020. "I'm just trolling: The role of normative beliefs in aggressive behaviour in online gaming", *Computers in Human Behavior, 102*, pp. 303-311.

M. Hindman, 2009. *The Myth of Digital Democracy*. Princeton University Press.

J. Hunsinger, L. Klastrup, & M. Allen, 2010. *International Handbook of Internet Research*. Springer.

I. Hutchby, 2001. "Technologies, texts and affordances", *Sociology: The Journal of the British Sociological Association, 35*(2), pp. 441-456.

E. A. Jane, 2014a. "'Back to the kitchen, cunt': speaking the unspeakable about online misogyny", *Continuum, 28*(4), pp. 558-570.

E. A. Jane, 2014b. "Beyond antifandom: Cheerleading, textual hate and new media ethics", *International Journal of Cultural Studies, 17*(2), pp. 175-190.

E. A. Jane, 2014c. "Your a Ugly, Whorish, Slut", *Feminist Media Studies, 14*(4), pp. 531-546.

E. A. Jane, 2015. "Flaming? What flaming? The pitfalls and potentials of researching online hostility", *Ethics and Information Technology, 17*(1), pp. 65–87.

D. Janssen & R. Kies, 2005. "Online Forums and Deliberative Democracy", *Acta Politica, 40*(3), pp. 317-335.

C. Jørgensen, 1998. "Public Debate – An Act of Hostility?", *An International Journal on Reasoning, 12*(4), pp. 431-443.

J. M. Kayany, 1998. "Contexts of uninhibited online behavior: Flaming in social newsgroups on usenet", *Journal of the American Society for Information Science, 49*(12), pp. 1135-1141.

H. Kennedy, 2006. "Beyond anonymity, or future directions for internet identity research", *New Media & Society, 8*(6), pp. 859-876.

S. Kiesler, J. Siegel & T. W. McGuire, 1984. "Social psychological aspects of computer-mediated communication", *American Psychologist, 39*(10), pp. 1123-1134.

S. Kiesler, D. Zubrow, A. M. Moses V. & Geller, 1985. "Affect in Computer-Meditated Communication: An Experiment in Synchronous Terminal-to-Terminal Discussion", *Human–Computer Interaction, 1*(1), pp. 77-104.

J. Kjeldsen, 2008. "Retoriske omstændigheder", *Rhetorica Scandinavica* (48), pp. 42-61.

C. Kock, 2009. "Choice is Not True or False: The Domain of Rhetorical Argumentation", *Argumentation, 23*(1), pp. 61-80.

C. Kock, 2018. "For deliberative disagreement: its venues, varieties and values", *Paradigmi, Rivista di critica filosofica, 3*, pp. 477-498.

J. Korenman & N. Wyatt, 1996. "Group Dynamics in an E-mail Forum". In: S. C. Herring (editor), *Computer-mediated Communication: Linguistic, Social and Cross-cultural Perspectives*, J. Benjamins, pp. 225–242.

L. M. Kruse, D. R. Norris & J. R. Flinchum, 2018. "Social Media as a Public Sphere? Politics on Social Media", *The Sociological Quarterly, 59*(1), pp. 62-84.

A. Kuntsman, 2007. "Belonging through violence: Flaming, erasure, and performativity in queer migrant community". In: K. O'Riordan & D. J. Phillips (editors.), *Queer online: Media technology and sexuality*, Peter Lang Publishing Inc, pp. 101–120.

P. G. Lange, 2007. "Commenting on comments: Investigating responses to antagonism on youtube". Paper presented at Society for Applied anthropology conference, Tampa, Florida.

N. Lapidot-Lefler & A. Barak, 2011. "Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition", *Computers in Human Behavior, 28*(2).

M. Lea, T. O'Shea, P. Fung & R. Spears, R, 1992. "'Flaming'" in Computer-mediated communication: Observations, Explanations, Implications". In: M. Lea (editor), *Contexts of Computer-Mediated Communication,* Harvester Wheatsheaf, pp. 89–112.

M. Lund, 2012. Provocative Style: The Gaarder Debate Example. In: C. Kock & L. S. Villadsen (editors.), *Rhetorical Citizenship and Public Deliberation*. The Pennsylvania State University Press, pp. 101-114.

N. B. McCormick & J. W. McCormick, 1992. "Computer friends and foes: Content of undergraduates' electronic mail", *Computers in Human Behavior, 8*(4), pp. 379-405.

A. McCosker, 2014. "Trolling as provocation: YouTube's agonistic publics", *Convergence, 20*(2), pp. 201-217.

C. R. Miller & D. Sheperd, 2004. "Blogging as Social Action: A Genre Analysis of the Weblog". In: L. Gurak, S. Antonijevic, L. Johnson, C. Ratcliff, & J. Reyman (editors), *Into the Blogosphere: Rhetoric, Community, and Culture of Weblogs*, http://blog.lib.umn.edu/blogosphere/.

E. Milne, 2010. *Letters, postcards, email—Technologies of presence*. Routledge.

P. J. Moor, A. Heuvelman & R. Verleur, 2010. "Flaming on YouTube". *Computers in Human Behavior, 26*(6), pp. 1536–1546.

C. Mouffe, 2005. *On the political*. Routledge.

NTB, 2020. "Frp-politiker misliker hets, tar ikke gjenvalg til Stortinget", *Aftenposten,* 11 April, https://www.aftenposten.no/norge/politikk/i/g7nKP1/frp-politiker-misliker-hets-tar-ikke-gjenvalg-til-stortinget [Accessed on 23 January 2020].

S. O'Riordan, J. Feller & T. Nagle, 2012. "Exploring the affordances of social network sites: An analysis of three networks". Paper presented at European Conference on Information Systems.

P. B. O'Sullivan & A. J. Flanagin, 2003. "Reconceptualizing 'flaming' and other problematic messages", *New Media & Society, 5*(1), pp. 69-94.

M. R. Parks & K. Floyd, 1996. "Making friends in cyberspace". *Journal of Computer-Mediated Communication, 1*(4).

W. Phillips, 2011. "LOLing at tragedy: Facebook trolls, memorial pages and resistance to grief online". *First Monday, 16*(12).

W. Phillips, 2013. "The House That Fox Built: Anonymous, Spectacle, and Cycles of Amplification", *Television & New Media, 14*(6), pp. 494-509.

W. Phillips, 2015. *This Is Why We Can't Have Nice Things Mapping the Relationship between Online Trolling and Mainstream Culture*. The MIT Press.

W. Phillips, 2019. "It Wasn't Just the Trolls: Early Internet Culture, 'Fun', and the Fires of Exclusionary Laughter", *Social Media + Society, 5*(3).

T. Postmes, R. Spears & M. Lea, 2000. "The formation of group norms in computer-mediated communication", *Human Communication Research, 26*(3), pp. 341-371.

P. Shachaf & N. Hara, 2010. "Beyond vandalism: Wikipedia trolls", *Journal of Information Science, 36*(3), pp. 357-370.

C. Shirky, 2008. *Here Comes Everybody*. Penguin Books.

R. Spears & M. Lea, 1992. "Social influence and the influence of the 'social' in computer-mediated communication". In: M. Lea (editors), *Contexts of computer-mediated communication*, Harvester Wheatsheaf, pp. 30–65.

J. Suler, 2004. "The online disinhibition effect", *Cyber Psychology & Behavior, 7*(3).

O. Tenenboim & A. A. Cohen, 2015. "What prompts users to click and comment: A longitudinal study of online news", *Journalism, 16*(2), pp. 198-217.

S. Thacker & M. D. Griffiths, 2012. "An exploratory study of trolling in online video gaming", *International Journal of Cyber Behavior, Psychology and Learning, 2*(4), pp. 17-33.

P. A. Thompsen, 1996. "What's fueling the flames in cyberspace? A social influence model", *Communication and cyberspace: Social interaction in an electronic environment, 2*, pp. 329-347.

P. A. Thompsen & D. A. Foulger, 1996. "Effects of pictographs and quoting on flaming in electronic mail", *Computers in Human Behavior, 12*(2), pp. 225-243.

K. Thorson, E. Vraga & N. Kligler-Vilenchik, 2015. "Don't push your opinions on me: Young people and political etiquette on Facebook". In: J. Hendricks & D. Schill (editors.), *Presidential campaigning and social media*, Oxford University Press, pp. 74–93.

E. Vatnøy, 2017. *The Rhetoric of Networked Publics. Studying Social Network Sites as Rhetorical Arenas for Political Talk.* Doctoral Thesis, University of Bergen.

A. Vromen, B. D. Loader, M. A. Xenos & F. Bailo, 2016. "Everyday Making through Facebook Engagement: Young Citizens' Political Interactions in Australia, the United Kingdom and the United States", *Political Studies, 64*(3), pp. 513-533.

S. S. Vrooman, 2002. "The art of invective: Performing identity in cyberspace", *New Media & Society, 4*(1), pp. 51-70.

J. Waldron, 2012. *The harm in hate speech*. Harvard University Press.

C. Wales, S. Cotterill & G. Smith, 2010. "Do citizens "deliberate" in on-line discussion forums? Preliminary findings from an Internet experiment", Paper presented at ECPR general conference, Potsdam, Germany.

H. Wang, 1996. "Flaming: More than a necessary evil for academic mailing lists", *The Electronic Journal of Communication, 6*(1).

B. Warnick, 1998. "Rhetorical Criticism of Public Discourse on the Internet: Theoretical Implications", *Rhetoric Society Quarterly, 28*(4), pp. 73-84.

B. Warnick, 2007. *Rhetoric online : persuasion and politics on the World Wide Web* (volume 12), Peter Lang.

S. F. Waterloo, S. E. Baumgartner, J. Peter & P. M. Valkenburg, 2018. "Norms of online expressions of emotion: Comparing Facebook, Twitter, Instagram, and WhatsApp", *New Media & Society, 20*(5), pp. 1813-1831.