SC|M
Studies in Communication and Media

FULL PAPER

Considering the Elaboration Likelihood Model for simulating hate and counter speech on Facebook

Potenziale des Elaboration Likelihood Model zur Simulation von Hass und Gegenrede auf Facebook

Carla Schieb & Mike Preuss

Carla Schieb (M.Sc.), Department of Communication, University of Münster, Bispinghof 9-14, 48143 Münster, Germany; Contact: carla.schieb(at)uni-muenster.de

Mike Preuss (Ass. Prof.), Leiden Institute of Advanced Computer Science, Universiteit Leiden, Snellius Building, Niels Bohrweg 1, 2333 CA Leiden, Netherlands; Contact: m.preuss(at) liacs.leidenuniv.nl

FULL PAPER

Considering the Elaboration Likelihood Model for simulating hate and counter speech on Facebook

Potenziale des Elaboration Likelihood Model zur Simulation von Hass und Gegenrede auf Facebook

Carla Schieb & Mike Preuss

Abstract: Counter speech is perceived and has also been advocated by social networks as a measure for delimiting the effects of hate speech. To facilitate estimating the efficiency of counter speech in freely accessible blackboard communication as employed by Facebook, we extend an existing simulation model by integrating Elaboration Likelihood Model (ELM) mechanisms. We model four different user groups (core, clowns, followers and counter speakers), each with a specific set of properties, namely need for cognition and involvement as ELM personal characteristics, and opinion, volatility and activity as borrowed from opinion formation models. We also add argument strength as important message characteristic. Our simulation experiments show that the updated model provides similar but much more detailed results: potentially temporal opinion changes via peripheral processing get visible. Furthermore, we give more evidence that the opinions of counter speakers shall not be too extreme in all cases but sometimes rather moderate in order to achieve maximum impact.

Keywords: Counter speech, hate speech, Elaboration Likelihood Model, Facebook, simulation

Zusammenfassung: Gegenrede wird als mögliche Maßnahme wahrgenommen und auch von sozialen Netzwerken empfohlen, um die Auswirkungen von Hassrede zu begrenzen. Um die Tauglichkeit von Gegenrede in frei zugänglichen Foren wie z. B. Facebookseiten einschätzen zu können, erweitern wir ein existierendes Simulationsmodell um Mechanismen des Elaboration Likelihood Model (ELM). Wir modellieren vier verschiedene Nutzergruppen (Kerngruppe, Trolle, Mitläufer, Gegenredner) mit jeweils unterschiedlichen Eigenschaften, insbesondere Need for Cognition und Involvement als persönliche Charakteristiken des ELM, außerdem Meinungsorientierung, Unbeständigkeit und Aktivität, die von Meinungsbildungsmodellen entlehnt sind. Wir ergänzen auch die Stärke eines Arguments als wichtige Eigenschaft einer Nachricht. Unsere Simulationsexperimente zeigen, dass das erweiterte Simulationsmodell ähnliche, aber detailliertere Ergebnisse liefert als das Ausgangsmodell. Eher temporäre Meinungsänderungen durch periphäre Verarbeitung werden jetzt sichtbar. Außerdem liefert das Modell Hinweise darauf, dass die von Gegenrednern vertretene Meinung nicht immer extrem, sondern unter bestimmten Bedingungen eher moderat sein sollte, um maximale Wirkung zu entfalten.

Schlagwörter: Gegenrede, Hassrede, Elaboration Likelihood Model, Facebook, Simulation

1. Introduction

The internet affords boundless, inexpensive, and ubiquitous communication, providing individuals with immediate information, enabling to share opinions, and bringing people together. There were high hopes in its diffusion in the late 1990s and early 2000s (Deuze, 1999; Elin & Davis, 2002; Shane, 2004). However, along with its benefits, one paradoxical effect was the noticeable rise of hateful speech and other antisocial activities in the form of websites, communities, postings, comments, pictures, and videos (Cammaerts, 2009; Citron & Norton, 2011; Erjavec & Kovačič, 2012; Gerstenfeld, Grant, & Chiang, 2003; Glaser, Dixit, & Green, 2002; Shepherd, Harvey, Jordan, Srauy & Miltner, 2015). Hate speech can be shared quickly via social networks and reaches large audiences spreading its toxic content (Awan, 2016; Benesch, 2014a; Gagliardone et al., 2016; Maynard & Benesch, 2016).

Crude and hateful content is found to be shared all over social media, for example *Twitter* (Chatzakou, Kourtellis, Blackburn, De Cristofaro, Stringhini, & Vakali 2017; Mondal, Silva, & Benevenuto, 2017; Groshek & Cutino, 2016; Burnap & Williams, 2015; Chen, Zhang, Chen, Xiang, & Zhou, 2015), *Facebook* (Hanzelka & Schmidt, 2017; Awan, 2016; Räsänen, Hawdon, Holkeri, Keipi, Näsi & Oksanen, 2016; Zerback & Fawzi, 2016), *4chan* (Hine et al., 2017; Bernstein, Monroy-Hernández, Harry, André, Panovich & Vargas, 2011), or *reddit* (Chandrasekharan, Pavalanathan, Srinivasan, Glynn, Eisenstein, & Gilbert, 2017; Mohan, Guha, Harris, Popowich, Schuster, & Priebe, 2017). Most social networking sites prohibit hate speech by their community standards or terms of service, e.g., verbal attacks and hatred based on people's race, ethnicity, national origin etc., which is closely related to scholarly definitions of hate speech (Benesch, 2012; Boyle, 2001; Delgado & Stefancic, 2014; Gagliardone, Gal, Alves, & Martinez, 2015; Gagliardone et al., 2016; Parekh, 2006). In reddit's case, however, volunteer moderators are in charge of detecting misconduct which includes manipulative actions and its community rules vaguely advise users not to insult others. Facebook's community standards state that hate speech is not prohibited per se, but is allowed under certain circumstances, such as expression of humor/satire, raising awareness for certain topics etc. Social network sites rely on counter speech as the means of choice. Users are emboldened to respect each other and to treat each other mindfully, their desire to discuss controversial topics is given an environment which leads to the experience of a sense of self-efficacy while counter-arguing. “When used wisely, counter speech may prove to be a very effective solution for harmful or threatening expression.” (Richards & Calvert, 2000). Counter speech is advocated to minimize the risks of violent acts (Benesch, 2014a) by encouraging audiences to take a stand against individuals who spread hate and mistrust or against platforms where hate speech is disseminated. As an example, the American right-wing news website “Daily Stormer” incited a hate campaign against Heather Heyer, the young woman killed in the car incident at the white supremacists rally in Charlottesville, New Jersey, on August 12, 2017. Public complaints led the website's hosting provider to cancel the contract and remove the “Daily Stormer” from their servers within a day.[1] One may see this as a different, but effective form of counter speech.

Both hate speech and counter speech are assumed to have persuasive effects on their audiences. There is substantial work on the immediate and long-term effects of hate speech which affects not only targets’ but also audience’s emotions, cognitions, and behaviors (Crowley, 2014; Gelber & McNamara, 2016; Soral, Bilewicz, & Winiewski, 2018). This is even more evident as hate speech containing overt calls for violence is deemed dangerous speech implicating imminent threat for targets of hate speech (Benesch, 2012). Much hope is set on counter speech to erase or at least minimize the harmful effects of hate speech. Its appeal is normative in nature but reliable empirical evidence is scarce, as there are only descriptive case studies. We aim to show that counter speech has a persuasive effect on audiences on social networking services by applying a widely known theory, the Elaboration Likelihood Model (ELM). The ELM is a dual process theory suggesting individuals use two different routes of message elaboration depending on their ability and motivation (Petty & Cacioppo, 1986). Individuals not able or not willing to reflect upon the provided information process messages cursorily by using heuristics or social cues while those who are interested and capable rely heavily on reasoning and arguments. While the former is said to have short-time effects on attitudes, the latter is considered to produce stable effects, especially when arguments are repeated. Studies applying the ELM as a theoretical framework show attitude changes primarily in experiments demonstrating causal relationships. This may not pose a greater problem for cursory processing of messages which yields short-dated attitude changes. However, evidence for stable attitude changes is weak as most studies are cross-sectional rather than longitudinal (Lee,2012).

A promising methodological approach arises from the field of computational social science. Agent-based models (Conte & Paolucci, 2014; Epstein, 1999; Waldherr, 2014) and computational simulations in general demonstrate complex interactions and effects in a defined system yielding “an optimal compromise between the model’s complexity and the complexity of the real world” (Voinea, 2016, p. xxii). Attitudes and attitude changes have been modeled in multiple computational simulations (Voinea, 2016, for a review) but to our knowledge only one research team has implemented the ELM in a computational simulation (Mosler, Schwarz, Ammann, & Gutscher, 2001). In Schieb and Preuss (2016), a simulation model has been established which examined the effect of counter speech as can be expected in the comments section on Facebook. The simulations have shown that, starting from the specific assumptions, counter speech has a very limited effect, except if counter speakers make up a reasonable, not too small proportion of the audience. However, the model was undertheorized as its assumptions relied solely on three defining properties, namely a user’s degree of activity, the valence of the shared opinion, and finally its volatility.

Our primary aim in this article is to evaluate the efficacy of counter speech by means of a computational model from the theoretical perspective of the ELM.

[1] https://www.theguardian.com/technology/2017/aug/14/anonymous-hackers-take-over-neo-nazi-website-daily-stormer-charlottesville-heather-heyer

https://doi.org/10.5771/2192-4007-2018-4-580 https://www.inlibra.com/de/agb - Open Access -

The main question is: Is it reasonable to encourage social media users to use counter speech against hate posts? We choose a simulation model for a number of reasons: First and foremost, reliable data is not available. Twitter data is not useful because we are aiming at blackboard discussions with a "group" character and a defined audience. Second, control conditions can easily be simulated. We can employ the model to answer questions such as: "What is the likely effect if the audience is twice as large?" Third, the operation is low-cost, unlike content analysis and ultimately, computing capacities allow for more sophisticated models.

The paper is structured as follows: At first, we delve into related work regarding hate speech and its proposed resolution, namely counter speech (Section 2). We then investigate the ELM (Section 3) and consider prior usage and how it may, at least in part, be used as basis for a quantitative model. Next, the simulation model is constructed (Section 4) as extension of the existing model established in Schieb & Preuss (2016). Furthermore, we report a number of simulations (Section 5), interpret the results and then conclude the paper.

2. Hate speech online and counter speech

Hate speech on the internet has recently captured a great deal of media attention, and it has become a key issue among legal institutions and policy-makers (Awan, 2016). Ensuing, there is also a growing body of research examining hate speech online. Though, it is striking that much research concentrates on many aspects of hate speech online, but without directly mentioning it, e.g., *(in-)civility/impoliteness* (Rösner, Winter, & Krämer, 2016; Alhabash, Baek, Cunningham, & Hagerstrom, 2015; Lange, 2014; Boyd, 2014; Megarry, 2014; Groshek & Cutino, 2016), *extremist/radical views* (Costello, Hawdon, Ratliff, & Grantham, 2016), *flaming* (Laineste, 2012), *trolling* (Herring, Job-Sluder, Scheckler, & Barab, 2002) or by subsuming different phenomena under the hate speech umbrella term, for example *cyberbullying* (Räsänen et al., 2016). Furthermore, a plethora of terms are used interchangeably, e.g., *cyber hate* (Burnap & Williams, 2015; Quandt & Festl, 2017; Douglas, McGarty, Bliuc, & Lala, 2005; Perry & Olsson, 2009), *E-bile* (Jane, 2014), *dangerous speech* (Maynard & Benesch, 2016), and *hate speech online/hate speech on the internet* being the most often used terms (Azriel, 2005; Cammaerts, 2009; Erjavec & Kovačič, 2012; Gagliardone et al., 2015; Leets, 2001; Nemes, 2002; Pollock, 2009; Shepherd, Harvey, Jordan. Srauy, & Miltner, 2015; Tsesis, 2001; Vollhardt, Coutin, Staub, Weiss, & Deflander, 2007). In this work, we follow George's (George, 2015) definition of hate speech which addresses all "forms of expression aimed at persecuting people by vilifying their racial, ethnic, or other identities. While the immediate target may be a single person or small group, the harm caused by hate speech can extend to entire communities by promoting discrimination and intolerance." (George, 2015, p. 305). George's definition implicates the persuasive effects of hate speech indicating potential harm not only to single representatives of targeted groups but the group as an integral whole. Assuming signs of threat or encouragement for physical violence, hate speech is regarded as dangerous speech (Gagliardone et al., 2016; Maynard & Benesch, 2016; Benesch, 2012) as it is a source of harm in general for those under attack (Waldron, 2012), when culminating in violent acts incited by hateful speech (Lawrence, 1990). Such violent hate crimes may erupt in the aftermath of certain key events, e.g., anti-Muslim hate crimes in response to the 9/11 terrorist attacks (King & Sutton, 2013). Furthermore, research shows that hate speech can deepen prejudice and stereotypes in a society (Citron & Norton, 2011) but also has a detrimental effect on the mental health and emotional well-being of targeted groups, especially on targeted individuals (Citron & Norton, 2011; Festl & Quandt, 2013; Benesch, 2014a).

Hate speech emotionalizes supporters, targets, as well as opponents because of its offensive language (Benesch, 2012; Parekh, 2006). Supporters are incited to shout out hateful words themselves or even perform violent acts to which they were called upon. Targets of hate speech experience psychological distress and other harms to their mental health (Gee, 2002; Delgado, 1993) while opponents (i.e., individuals who oppose hate speech but are not primarily targeted) may be feeling sensations of rage and anger or slight frustration. Reactions are expectable for these groups of people but the large majority of bystanders who have not yet formed strong attitudes towards the issues at hand are the focus in our paper. Will they adhere to hate speech or will counter speech persuade them and to what extent?

Counter speech as hate speech's antagonist is mostly a factual and objective argumentation strategy aiming to debunk hate speech and to strengthen the position of targets and opponents of hate speech by providing them with further arguments (Richards & Calvert, 2008). In addition, counter speech encompasses also the use of memes (Benesch, 2014), empathetic responses (ibid.) or even humorous and sarcastic responses of journalists in online discussion forums (Ziegele & Jost, 2016).

Whether it is hateful and inciting comments on online news web sites (Erjavec & Kovačič, 2012), on SNS such as Facebook or Twitter (Burnap & Williams, 2015), US-American internet content providers are free to choose how they respond to hate content. In essence, they may choose (1) inaction, (2) deletion of improper and hateful speech, (3) education and promotion of respectful conduct, or (4) addressing hate speech with counter speech (Citron & Norton, 2011). *Inaction* refers to simply ignoring hate content or establishing weak rules as in reddit's case. However, inaction can lead to greater harm to targets of hate speech and may demonstrate users that content providers do not take victims of hate speech seriously. Citron and Norton rank the removal of hateful speech as the most powerful tool at disposal (Citron & Norton, 2011). *Deletion* includes not only the removal of offensive, hateful content but also blocking users or shutting down their accounts. The latter options are chosen by content providers especially in case of violent threats towards individuals or certain social groups leaving criminal prosecution untouched. As with Germany's Network Enforcement Act this could lead to overblocking to prevent heavy fines. The third applicable option is *education*. Content providers could play an active role in promoting respectful behavior and thus informing about the harms resulting from hate speech. Furthermore, they could make their actions towards hate speech public, exposing their motives and hence taking a stand against hate speech. (Citron & Norton, 2011). *Counter speech* by online content providers themselves is rare but occurs from time to time. More often, counter speech is performed by users themselves and is meant to encourage users to understand and tolerate diverse opinions.

Counter speech is regarded as the most important response to hate speech, in fact as "constitutionally preferred" (Benesch, 2014a). Within the meaning of the First Amendment it is regarded as beneficial if "bad" speech is met with more speech, i.e., counter speech (Abdelkader, 2014; Richards & Calvert, 2000). Scholarly definitions of the term are scarce, rather some vague examples serve for clarification (Richards & Calvert, 2000; Benesch, 2014a; Henry, 2009). "Counterspeech is a common, crowd-sourced response to extremism or hateful content. Extreme posts are often met with disagreement, derision, and counter-campaigns" (Bartlett & Krasodomski-Jones, 2015). We define counter speech as all communicative actions aimed at refuting hate speech through thoughtful and cogent reasons, and true and fact-bound arguments. Such communicative actions can be memes such as the Panzagar (flower speech) meme of Burmese blogger Nay Phone Latt (Benesch, 2014b), the billboard of citizens in Missouri to respond to the Ku Klux Klan (Richards & Calvert, 2000), or information spread in online hate groups by the Southern Poverty Law Center (Henry, 2009) and many other means to fight hate speech.

Academic work on counter speech is descriptive in nature and tackles the subject matter merely in terms of successful case studies (see Ziegele, Jost, Bormann, & Heinbach and Leonhard, Rueß, Obermaier, Reineman in this issue for two striking exemptions). Research on counter speech lacks analytical rigor and its effects are not systematically connected to more sophisticated approaches such as works on argument strength (Stephenson, Benoit, & Tschida, 2001). Counter speech is thoughtful reasoning by definition and as such is expected to show similar outcomes as strong arguments as applied in Petty and Cacioppo's (1986) Elaboration Likelihood Model. Our aim is to move beyond and to turn towards an analytically more sophisticated approach which is able to identify the potential counter speech may have in an instigative environment.

3. The Elaboration Likelihood Model

Both hate speech and counter speech are regarded as persuasive messages capable to influence emotions, attitudes, and even the behavior of bystanders depending on individual and message characteristics. While hate speech effects have been investigated repeatedly, counter speech effects have been shown mostly in juridical articles discussing particular instances (e.g., Abdelkader, 2014; Bartlett & Krasodomski-Jones, 2015) but systematic investigations are scarce (Schmitt, Rieger, Rutkowski, & Ernst, 2018). Research suggests that sound and truthful arguments are considered to be strong, i.e., being effective in attitude change (Cacioppo, Petty, & Morris, 1983; Petty & Wegener, 1998; Stephenson et al., 2001). As outlined above counter speech is regarded to have properties (truthfulness, validity) which can be applied to the strong arguments variable within the ELM (Petty & Cacioppo, 1986).

The ELM is an approach often used when shifts or changes of attitudes are under investigation. It has its origins in consumer research (Petty, Cacioppo, & Schumann, 1983; Petty & Cacioppo, 1986) and has been applied various times in this field (Cheng & Loi, 2014; Malthouse, Calder, Kim, & Vandenbosch, 2016; Orizio, et al., 2010; SanJosé-Cabezudo, Gutiérrez-Arranz, & Gutiérrez-Cillán, 2009). Although it has generated much academic debate over time (Areni, 2003; Bitner & Obermiller, 1985; Johnson & Eagly, 1989; Kitchen, Kerr, Schultz, McColl, & Pals 2014; Stiff, 1986), the ELM has wide appeal to date because it managed to clarify conceptual inaccuracies (Briñol & Petty, 2012; Schumann, Kotowski, Ahn, & Haugtvedt, 2012). Furthermore, the rationale behind the ELM is clear-cut and its straightforwardness has attracted scholars of subjects such as *propaganda research* (Müller, van Zoonen, & Hirzalla, 2014), *health communication* (Withers & Wertheim, 2004; Withers, Twigg, Wertheim, & Paxton, 2002), *aggression* (Douglas, Kiewitz, Martinko, Harvey, Kim, & Chun 2008; Foubert & Perry, 2007), and *hate speech* (Lee & Leets, 2002).

The ELM is conceptualized as a dual-processing model claiming that attitude shifts occur through a peripheral or a central route of elaboration (Petty & Cacioppo, 1986). The cognitive endeavor applied when a persuasive message is being processed does not represent distinct categories ("heavy thinking," "no thinking at all") but is rather conceptualized as a continuum with two ideal outcomes, namely the two routes of elaboration. The *peripheral route* produces unstable and ephemeral attitudes; individuals rely mainly on cursory cues, whereas the *central route* implies a mental involvement "a person's careful and thoughtful consideration of the true merits of the information presented" (Petty & Cacioppo, 1986, p.125), resulting in rather stable attitudes. Chaiken's heuristic-systematic model (HSM) (Chaiken, 1980, 1987) is similar to the ELM in that it proposes two different ways of information processing, namely one in which individuals rely on heuristics and another which draws from systematic evaluation of available information. Nevertheless, there are significant differences between the ELM and the HSM (Johnson, Maio, & Smith-McLallen, 2005) which we do not discuss here. The main reason for choosing the ELM over the HSM is that to our knowledge there is neither an HSM computational model nor has HSM been applied to study hate speech or counter speech effects. Both is true concerning the ELM: Mosler et al. (2001) used the ELM to implement a computational simulation. Furthermore, Lee and Leets' 2002 study applied the ELM to examine persuasive story-telling effects of hate speech. We are interested in counter speech effects and use the ELM framework to model a computational simulation.

A certain configuration of several variables, such as characteristics of the target person herself, the message, and the source (e.g., her credibility, attractiveness) helps to predict the likelihood of a (persuasive) message. Beginning with individual variables of the target person, need for cognition has been found to be the most important (Haugtvedt, Petty, & Cacioppo, 1992; Cacioppo, Petty, Feinstein, & Jarvis, 1996; Elias & Loomis, 2002). *Need for cognition* refers to the preference to enjoy cognitive challenges. Individuals with a high need for cognition enjoy deepening their knowledge and thus tend to process given information via the central route of elaboration, whereas individuals with low need for cognition are not interested to consider the pros and cons of arguments and thus rely mainly on peripheral cues. *Involvement* is a moderator variable, mostly defined as issue involvement with some degree of "personal relevance or consequence" (Petty et al., 1983, p. 136). Individuals who are highly involved process given information via the central route, while those less involved will lead to peripheral processing. *Argument strength* on the other hand is a characteristic of the message and it can have a persuasive force on individuals capable and motivated enough to process the presented information (the central route of elaboration is chosen). Then again, inattentive, distracted individuals cannot tell the difference between strong and weak arguments, and thus they are more attracted to weak arguments and/or peripheral cues. However, as research in the field of cognitive psychology shows, the effect of motivated reasoning does not hinder people to cling to false beliefs, although they are presented with strong arguments (Kunda, 1990, for an overview, also Lodge & Taber, 2000). In our present study, we proceed on the assumption that individuals are able and willing to process hate speech or counter speech, depending on the configuration in our simulation model (see following section). The basic reasoning in integrating the ELM's related factors into the simulation is that we use need for cognition and involvement to decide if the peripheral route or the central route is employed: individuals with high need for cognition and involvement will process via the central route, all others via the peripheral route. Argument strength is used to compute the amount of opinion shift that results from comments on social media. A special case results if a weak argument is processed via the central route: the argument is simply dismissed and does not provoke any opinion shift. These mechanisms are controlled via parameters (i.e., what exactly does it mean if an argument is strong in an interval from 0 to 1?) which are chosen assuming maximum realism.

4. Simulation model

Given that compared to the natural sciences, computer simulation is much less prevalent in social sciences, one may ask what we can actually achieve by that in the context of opinion shifts in blackboard communication. Generally, simulating means to rebuild the essential parts of a complex system as a computer program that can then be used to explore the (simplified) system behavior under different starting conditions, see how it reacts to unexpected events, or look into the (possible) future. Albeit simplified, the simulation system is still much too complex to predict its outcome by means of a thought experiment, usually because there are too many actors and it is unclear how their individual actions add up to a specific system behavior.

As it is often difficult to show that one has captured all relevant mechanisms of the modeled system as well as selected matching starting parameters, there is no guarantee that in reality, the modeled system will behave similar to the simulation model. However, as many simulations include randomized mechanisms or starting conditions, they can be run repeatedly and thereby provide a distribution of possible results. The computer model also forces us to explicitly provide concrete values and equations, which requires that one has to postulate relevant mechanisms and estimate values that can afterwards be checked by comparing model outcome to real-world developments or be adapted by integrating new knowledge.

To investigate the effect of counter speech on a Facebook or similar small ad hoc audience online media page, we have suggested a simple simulation model in Schieb & Preuss (2016). In this work, we integrate the main components of the ELM into our simulation. This makes the model somewhat more complex, but at the same time we also get much more detailed data, more specifically on presumably permanent and nonpermanent opinion changes due to the different routes of persuasion the ELM postulates. Note that the ELM (Petty & Cacioppo, 1986) is an explanatory model, not a numeric model. Therefore, it requires a certain amount of interpretation and concretion to select the components we want to add to our simulation model, and to set the formulas and required threshold values right, thereby preserving the spirit of the ELM as closely as possible. Among the many personal characteristics that influence persuasion according to the ELM, we only add two to our model: need for cognition and involvement. We employ these in order to decide how individuals process the arguments they are confronted with, either using the central or the peripheral route (see Section 3). As the decision also depends on the strength of the argument, we also model this as numerical factor between 0 (very weak argument) and 1 (very strong argument).

Apart from these new factors, we keep the overall architecture of the model largely intact: each individual post (or read) on a (Facebook) blackboard is modeled as an action that can potentially change the opinion of all board visitors, and all posts express an opinion on a one-dimensional opinion scale ranging from –1 to 1. W.l.o.g., –1 stands for extreme hate speech throughout the paper, and 1 for extreme counter speech. Of course, the same model can also be used in a different setting (e.g., for discussing a movie), then –1 could mean strongly against, and 1 strongly in favor. As the effect of likes in our previous study had been only marginal and we want to concentrate on the effect of adding ELM based mechanisms, likes are completely removed from the simulation.

We know the model can only help in identifying trends, the numerical results cannot be directly transferred to implementable policies. However, we can ask "what if"-questions and find out what kind of knowledge is most urgently needed, because certain parameters or conditions have more influence than others. Furthermore, we want to obtain a general idea of how much counter speech is needed to balance the *leading opinion* or even revert it, and what are the important factors for the effect of counter speech. The contents of a post are not modeled, we only look at the influence exerted on participants of such a blackboard at a given point in time (i.e., the comments section). Therefore, the updated simulation model is based on the following assumptions:

- All posts are concerned with only one general area of opinion, that is, participants do not discuss completely different matters, but focus on a single, possibly very general topic. As an example, this could be the immigration of refugees into Germany. One consequence of this assumption is that the involvement of a specific individual is assumed to be constant during the simulation.

- Opinions of participants are well reflected in their posts, so that they can be recognized as expressions of a specific opinion by the audience. Posters may use language skills as irony or sarcasm, but readers are still able to determine which opinion is expressed (positive or negative in different strengths, or neutral).
- Most participants, except the ones with extreme opinions, can be influenced by counter speech, and they change their opinions only gradually as a reaction to the posts they see, or respectively, the opinions that are expressed by these posts. This change is at least in principle possible in both directions (positive, negative).
- Individuals pay attention to the blackboard discussion on different levels. There are some with a high need for cognition that thoroughly evaluate the arguments and some with a lower need for cognition that are rather reacting to peripheral cues. We presume that need for cognition is a characteristic of an individual and stays constant for the whole time span of a simulation.

Table 1. Property values for the different modeled groups, * means Gaussian distribution with mean 0.5 and standard deviation 0.15, other ranges stand for uniform distributions

Property/interval	Core	Clowns	Followers	Counter Speaker
Opinion	–1	[–1, 0]	[–1, 0]	1/0.5
Volatility	0	[0, 1]/0	[0, 1]	0
Activity	1	1	[0, 1]	1
Need for cognition	1	1	[0, 1]*	1
Involvement	1	1	[0, 1]	1
Default group distribution	10%	5%	≤ 85%	var

We call the fixed group of participants that is directly or indirectly involved in the discussion on a specific board (e.g., Facebook page) of an SNS at a given time the *audience*. Note that this does not even potentially encompass all Facebook users, but only the ones who visit a specific page within a certain time interval. The audience consists of participants in 2 factions, namely supporters of the original post that is assumed to be hate speech, and counter speakers. Whereas the latter group is homogeneous, the supporters come in three types: *core*, *clowns*, and *followers*. Members of the core have extreme opinions and no volatility, that is, they cannot be influenced.

Clowns follow the haters, have less extreme opinions and a high activity. This group is related to people known as *trolls* in other network contexts (Buckels, Trapnell, & Paulhus, 2014). Followers are much easier to influence than the core, but have a lower activity. Counter speakers are the core's antagonists, which means that they also do have extreme opinions, but at the other side of the allowed interval, and they are also highly active and cannot be influenced. The intervals for all groups are given in table 1.

Consequently, the blackboard we model is not at all neutral concerning the average of the user opinions. We assume a situation where the board has been

https://doi.org/10.5771/2192-4007-2018-4-580 https://www.inlibra.com/de/agb - Open Access -

primed in a certain direction and mostly attracts users with a similar, but mostly less extreme opinion. This is consistent with the confirmation bias theory which states that individuals pay attention primarily to information confirming one's own viewpoint (Wason, 1968).

Our general approach to simulate a mutual influencing process is related to agent-based modeling (see Heath, Hill, & Ciarallo, 2009 for an example-based survey), only that in our model, an agent is little more than a container for five numbers that represent its defining properties. It is thus similar to the approach pursued with opinion formation models (Watts & Dodds, 2007), only that we ignore the network component here and presume that every participant in the audience is able to see every post on a specific (modeled) Facebook page:

- Opinion $o \in [-1, 1]$, where -1 stands for the one extreme (in our context a hater), and 1 for the other extreme
- Volatility $v \in [0, 1]$, where 0 means that the opinion of the participant is not mutable at all, and 1 means that it is very easily influenced, and
- Activity $a \in [0, 1]$, which corresponds to the probability of a participant to actively take part in a discussion.
- Need for cognition $nc \in [0, 1]$, represents the general motivation of the participant to use the central route (thoughtfully consider arguments) for processing new information, 0 is very low, and 1 very high.
- Involvement $inv \in [0, 1]$, where 0 means no involvement at all (the participant is not interested in the topic), and 1 stands for very high involvement.

A participant and his or her behavior during the simulation is completely defined by the quintuple $p = (p_o, p_v, p_a, p_{nc}, p_{inv})$. To keep our simulation model simple, we assume the following influence process:

- Every participant in the audience can see posts and is influenced by them if the own volatility is > 0.
- Every participant can choose to act via writing an own post or do nothing. The probability of posting is controlled by the activity parameter pa.
- Every participant is allowed to perform one action only per iteration. That is, before a participant can act again, the current iteration including all participants must be finished.
- The potential of an influence change is generally higher if the difference in opinions between two participants is large, up to a maximum when the difference is exactly 1. For larger differences, we assume that the potential shrinks again, according to the triangular shaped potential of influence as utilized in Schieb & Preuss (2016). This simply reflects that humans with similar opinions cannot mutually influence each other much towards a different opinion. That does not necessarily mean that the resulting opinion change is always large. A participant's volatility acts as a filter here: low volatility reduces the change severely, high volatility enables it.

According to the given assumptions, we can handle the influencing process in a sequential manner, by computing the influence exerted by a single post on the rest of the audience. A single interaction is characterized by equation (1), using p as the posting participant, r as the receiving participant (p_O and r_O are the opinions of the poster and the receiver, respectively, and r_V is the volatility of the receiver), and D as damping factor that reduces the range of possible opinion changes within one interaction. This factor is set to $D = 0.1$ per default, but its importance for the simulation is limited because it slows down or speeds up all interactions at the same rate.

$$r_O := r_O + (1 - |1 - (p_O - r_O)|) * r_V * D * r_{nc} * r_{inv} * a_s \qquad (1)$$

The $(1 - |1 - (p_O - r_O)|)$ multiplier provides us with a triangular influence shape as discussed above. It appears as more realistic as the linear influence shape first assumed in (Schieb & Preuss, 2016). The triangular shape means that the influence is maximal when the difference of opinions between two simulated persons is exactly 1, which is half of the possible spectrum (–1 to 1). Are the opinions further apart, the influence shrinks again and is zero for the two extreme positions. Any other specific influence shape may be included by simply replacing the multiplier with another term.

As described above for participant p, the variables r_{nc} and r_{inv} represent need for cognition and involvement of the recipient, respectively. Argument strength is expressed as a_s, with possible values between (including) 0 and 1. In the following, we presume that the blackboard conversation that is modeled always starts with a strong argument, and if there is at least one counter speaker, a strong counter argument (both with argument strength of 1).

The overall simulation method is given by the pseudo-code algorithm 1 (table 2). “Perform post” in algorithm 1 is done by applying equation (1) sequentially on the whole audience, with the current participant as poster p. For each recipient, we have to decide whether the central or peripheral route are used to process the post. According to the general idea of the ELM, we choose the central route if need for cognition and involvement of the recipient are high. Setting a concrete threshold is somewhat arbitrary, as the ELM itself does not define what high and low are in numbers.

Without further knowledge available, we choose 0.5 as threshold in both cases, that is, for any (symmetrically) randomly initialized audience with need for cognition and involvement values between 0 and 1, about one quarter of the participants will utilize the central route because both values are at least 0.5. If the central route is taken, the argument strength is vital: if the argument is weak, it is simply dismissed by the recipient and no opinion change occurs. If it is strong (at least 0.5 on a scale between 0 and 1), the opinion of the recipient is changed via equation (1). In case the peripheral route is used, the opinion change always happens via equation (1), without precondition. All described settings and constant values have been chosen to represent the overall principles of the ELM, however, the exact values are of course debatable.

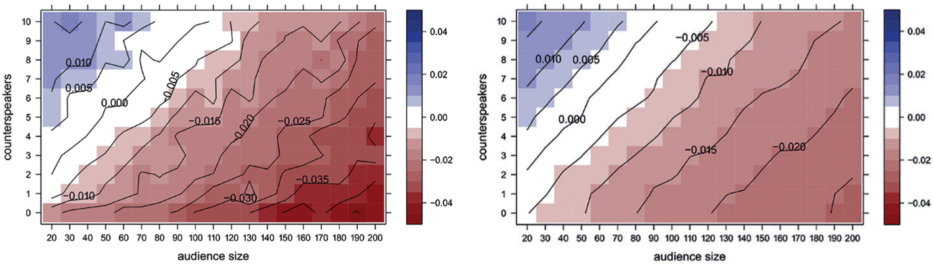
Table 2. Algorithm 1: Influence model

```
1   create initial *audience* according to values in table 1 ;
2   mark all *participants* ∈ *audience* as “ready”;
    *(this starts the discussion)*
3   select *supporter* with extreme opinion, perform post, mark as “done”;
4   **if** #*counter speakers* > 0 **then**
5       select *counter speaker*, perform post, mark as “done”;
6   **while** *participants* ∈ *audience* that are “ready” **do**
7       randomly select one of these as $p$;
8       **if** random number ∈ [0, 1] < $p_a$ **then**
            *(the participant “chooses” to get active)*
9           perform post;
10      mark $p$ as “done”;
        *(we can do more than one iteration by starting again)*
11  **if** !*termination* **then**
12  goto step 2
```

To enable the analysis of differences between central and peripheral processing later on, we also store the amount of opinion changes via each route as a variable for each member of the audience.

Note that in contrast to many other opinion formation models, we do not strive for a discrete state (as necessary for decisions, e.g., in an election context), but the participants may end up with gradually different opinions distributed over the whole possible interval [–1, 1].

Figure 1. No ELM (simple model, left) and ELM extended simulation results (right, 209 simulations with 50 repeats each) with audience size up to 200 and 0 to 10 counter speakers. At the 0 contour line, both influences cancel each other out on average, negative numbers (red) mean an overall shift towards the original (hater) post, positive numbers (blue) for a shift towards the opinion of the counter speakers.



5. Experimental analysis

Our simulation model is highly parametrizable, such that many different scenarios may be investigated. Our experiments can therefore not be comprehen-

https://doi.org/10.5771/2192-4007-2018-4-580 https://www.inlibra.com/de/agb - Open Access -

sive. Instead, we attempt to provide answers to the following two general questions:

- Does the ELM extended model provide similar results as the simpler model of (Schieb & Preuss, 2016)? What is the added value of the ELM?
- What are the differences between realistic use cases, e.g., counter speech on a hater forum, and a neutral blackboard?

5.1 ELM validation experiment

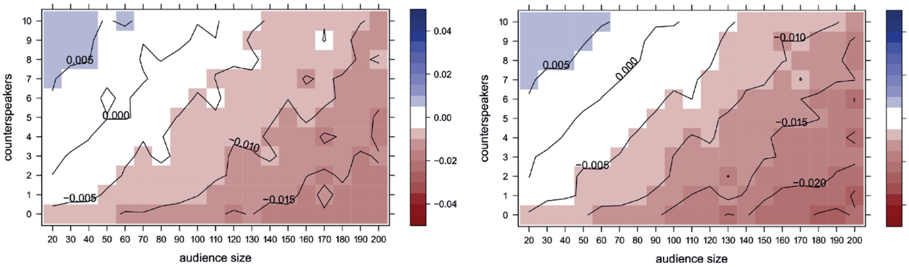
In our first experiment, we investigate if and how the obtained simulation results differ from each other if we compare the simple to the ELM extended simulation. Our use case is a blackboard (i.e., Facebook page) that is already biased towards the negative (hater) side and is entered by a group of counter speakers who attempt to persuade the audience towards the opposite direction. We presume that there are several opinion leaders (core) active on the page that approximately represents 10 % of the audience. As the page is biased, the rest of the audience (followers, clowns) also has an opinion ranging from 0 (neutral) to −1 (extreme hate), uniformly distributed.

Research question: Does the addition of ELM based mechanisms into the simulation model lead to significant changes and/or new insights?

Pre-experimental planning: We stay with 50 repeats per configuration as in Schieb & Preuss (2016) as a compromise between accuracy and computation time. Actually, the variance seems to be slightly larger for the ELM extended setting.

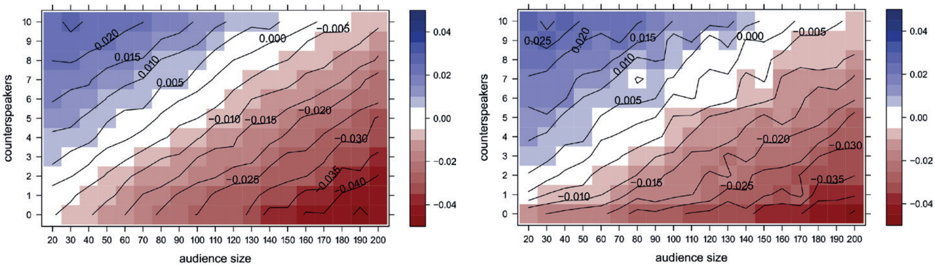
Setup: We run several simulations, varying the two parameters #counter speakers over #supporters of original post[2], for a fraction of core (haters) of 10 %, the remaining parameters are provided in table 1. The opposition opinion is set to 1. As result, we measure the average shift in opinions of the whole audience per configuration.

Figure 2. ELM results, same configuration as Figure 1, opinion shifts due to central route processing (left) and peripheral route processing (right).



[2] # stands for: „number of“

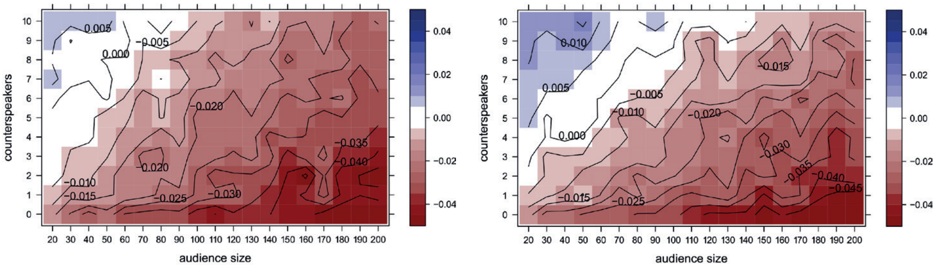
Figure 3. No ELM (left) vs. ELM (right) results, same configuration as Figure 1, but with an opposition opinion of 0.5.



Results: Figure 1 shows the simple model result for opposition opinion 1, and the corresponding result for the ELM. Figure 2 shows detailed views of the latter, including only central or peripheral processing, respectively. Figure 3 compares the described configurations above (ELM and no ELM) for an opposition opinion of 0.5. Note that each square in the figures corresponds to the average value of 50 independent runs.

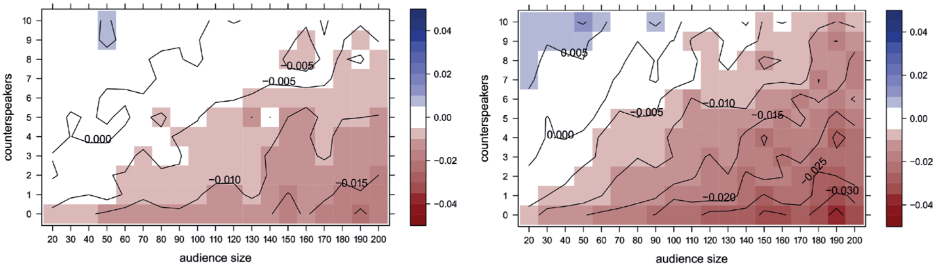
Observations: The overall opinion shift for the ELM and the simple model seem similar for both opposition opinion settings. However, the former seems to have more noise. The detailed ELM view shows that central route and peripheral route processing are comparably strong. Considering the influence exerted by the counter speakers for the 2 different opposition opinion settings, it seems that an opinion of 0.5 strengthens the counter speech effect considerably.

Figure 4. ELM-based simulation of a neutral blackboard, counter speech with moderate opinion (left) vs. counter speech with extreme opinion (right).



Discussion: For simulating the same situation with and without ELM extension, the model returns very similar results. The higher fluctuation of the ELM-based model is probably due to many more randomly distributed variables, as every single person is now by five instead of three properties. We also see that for approaching a strongly biased blackboard with counter speech, it makes more sense to use a moderate opinion position. This follows from the triangular shaped potential of influence as for too different opinions, there is only a smaller potential of influence. Interestingly, the quite small group within the audience that uses central processing is responsible for an overall opinion shift as the approximately three times larger group that uses the peripheral route only. Of course, the relations will change if the fraction of people who employ central processing grows or shrinks. However, it is remarkable to see that the exerted influence on most people is rather temporary.

Figure 5. ELM-based simulation of a neutral blackboard, extreme counter speech opinion, average opinion shift due to central processing (left) vs. opinion shift due to peripheral processing (right).



5.2 Comparing Differently Biased Blackboards

Research Question: Is the counter speech effect significantly different for different types of blackboard bias?

Pre-experimental planning: After some experimentation, we fix the parameters of the opinion distribution for the followers to mean 0 and standard deviation 0.5. This entails that most opinions are centered around neutral, but there is a considerable number of much more extreme opinions.

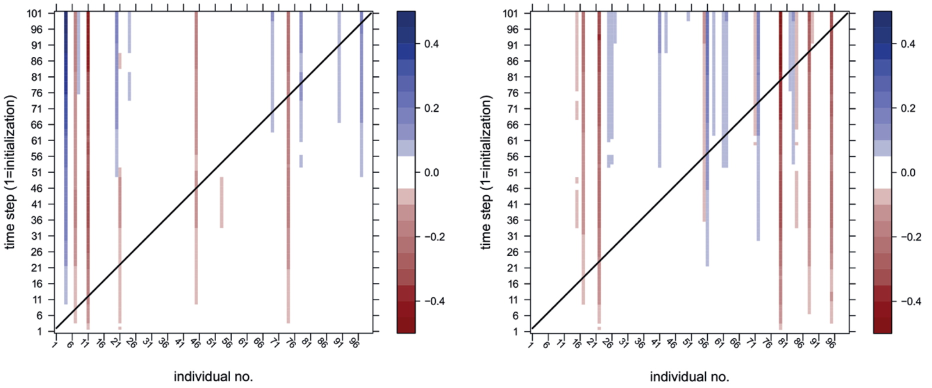
Setup: We run a similar setting as in the first experiment, but with the above described Gaussian follower distribution and counter speaker opinions of 0.5 and 1.0. Specifically, we still assume that a fraction of around 10 % of the audience consists of haters. The opinions of the clowns (5 % of the audience) are changed towards a uniform distribution over the whole opinion interval (between –1 and 1). As we realized that this also means that clowns can have stronger shifts in opinions (their opinion differences to haters can be much larger now), which may obscure the opinion shift of the rest of the audience, we have set their volatility to zero. This means they cannot change their opinion during the experiment.

Results: Figure 4 shows the overall opinion shifts for both counter speaker opinion stances. Figure 5 details the extreme counter speech simulation by showing the separate central processing and peripheral processing fraction of the opinion shift. In Figure 6 we see an example of the individual opinion change processes over the course of one simulation (note that only one run is shown such that different runs may have slightly different outcomes).

Observations: Comparing the two pictures of Figure 4 shows that in the neutral scenario, the overall influence shift towards the haters is stronger if the counter speakers have moderate instead of extreme positions. As for the previous simulation experiment, opinion changes via use of the peripheral route appear to be stronger than for central route processing, the difference is more emphasized than before. The in-depth look into one simulation provided in Figure 6 shows that only a small number of individuals change their opinion at all, much less due to central route processing.

Discussion: Interestingly, we see that (under our assumptions) on a neutral board, it seems to be even harder to balance hate speech with counter speech, and it makes more sense to take an extreme stance instead of a moderate one as for the first experiment. As the amount of noise is much stronger in the neutral board scenario, this result only applies on average and it is much less predictable what happens in one individual case, even if the start parameters are known. The look into one simulation shows that some individuals change their opinions gradually over time, especially in the peripheral processing case, as opposed to the central processing case, where fewer and stronger changes are visible. We presume that this is well in accordance with the ELM and appears to be realistic. Of course, this also means that the largest part of the opinion shifts is of rather temporary nature.

Figure 6. ELM-based single sample simulation of a neutral blackboard, 10 counter speakers with extreme opinions, audience size 100. Individual opinion shift due to central processing (left) vs. opinion shift due to peripheral processing (right) over time (from bottom to top. The diagonal line indicates the point in time when the individual acts).



6. Conclusion

We have discussed the current state of research for hate and counter speech and suggested to employ the Elaboration Likelihood Model (ELM) to better understand discussions on social media. Our simulation model attempts to provide the means to look deeper into the interaction schemes of such ad-hoc discussions and is highly configurable such that any specific situation seen in the real world may be replicated and maybe tested for the effect of parameter changes ("what if" questions).

Our simulation experiments examine a small fraction of what is possible. However, we have obtained some interesting insights:

- The results for the simple and our ELM extended model are relatively similar on the level of detail the simple model allows to compute. This is necessary, as we want to get more insight into the processes we model, but not completely different results. Nevertheless, the ELM based simulations unfortunately possess considerably more variance, which is of course due to the much higher complexity (e.g., every member of the audience is described with 5 properties instead of 3).
- The results for the simple and our ELM extended model are relatively similar on the level of detail the simple model allows to compute. This is necessary, as we want to get more insight into the processes we model, but not completely different results. Nevertheless, the ELM based simulations unfortunately possess considerably more variance, which is of course due to the much higher complexity (e.g., every member of the audience is described with 5 properties instead of 3).
- Depending on the situation, counter speech should not take too extreme positions. If a board that is primed into a certain direction is targeted, moderate counter speech works better than extreme positions (0.5 better than 1.0) because too extreme arguments are simply dismissed by a large part of the audience as the difference in opinion is too large. On a neutral board, the situation is the opposite: the more extreme the counter position, the better.
- More than half of the observed opinion shift is produced via peripheral route processing and presumably not durable, with our parameter settings this affects around three quarters of the audience.
- The first speaker always has an advantage. If hate speech is spread on the blackboard, this already moves the average opinion on the board, and this effect is passed on by later speakers. Counter speakers are in the more difficult situation.
- From this it follows that counter speech works best if it is organized, or at least conducted in groups. A single speaker will have difficulties to balance the influence exerted by haters in any case. Additionally, counter speakers ought to be quick. The longer it takes until they act, the more time the malicious opinion shifts have to spread.
- Overall, the provided results and conclusions justify the establishment of a relatively complex model for simulating hate and counter speech and to make it even more complex by adding effects postulated by the ELM. However, our simulations also show that there is considerable uncertainty in predictions obtained by means of this model. The results rather hold on average, but not necessarily for any single instance of a blackboard communication.

What do these results mean for real-world situations in which single persons or groups consider using counter speech against hate speech? According to our model, the counter speakers should know about the average opinion levels in the corresponding forum. If the audience is rather neutral, extreme positions make sense. If the audience is already strongly biased, counter speech should not be too extreme because otherwise it has much less influence on the opinions of the audience. Also, it is of course an advantage to act early, before the audience becomes too biased, and not to act alone but in groups.

An even deeper look into the effects, also on the level of single interactions, may be an interesting avenue for further research, but this would greatly benefit from a better empirical foundation. We have made several assumptions that can be questioned, but we have done so in absence of useful data. However, as soon as this data becomes available, it can easily be integrated into our model. Empirical insights that would be especially valuable could provide realistic numbers for shares of haters (core) and clowns (trolls, here we have at least one study providing an estimated value) in specific audiences. Also, the amplitude of opinion shifts can be adapted via the damping factor, and provided that for any specific case, this amplitude can be estimated, the simulation can be tuned to this case by lowering or increasing the factor. Three more important parameters are the thresholds for argument strength (for deciding if an argument is processed at all via the central route), involvement and need for cognition. The latter two determine which route is used for processing and requiring higher values for central processing would lead to a larger fraction of peripheral processing. In any case, the outcome of the simulation will always be one possible future, rather a trend than a result. Several simulation runs should be performed to generate meaningful statistics.

Nevertheless, we would like to encourage other researchers to use this type of simulation and to challenge our model with new and interesting setups that have not been taken into account yet. One of these possible extensions is the continuation of the simulation in terms of several iterations. Visitors of a blackboard forum as Facebook may spend longer on the discussion if it challenging or interesting, and post more than once. It shall be interesting to see how these alterations (possibly) change the big picture.

References

Abdelkader, E. (2014). Savagery in the subways: Anti-muslim ads, the first amendment, and the efficacy of counterspeech. *Asian American Law Journal*, *21*.

Alhabash, S., Baek, J.-H., Cunningham, C., & Hagerstrom, A. (2015). To comment or not to comment? How virality, arousal level, and commenting behavior on Youtube videos affect civic behavioral intentions. *Computers in Human Behavior*, *51*, 520–531.

Areni, C. S. (2003). The effects of structural and grammatical variables on persuasion: An elaboration likelihood model perspective. *Psychology & Marketing*, *20*(4), 349–375.

Awan, I. (2016). Islamophobia on social media: A qualitative analysis of the Facebook's walls of hate. *International Journal of Cyber Criminology*, *10*(1), 1–20.

Azriel, J. (2005). The internet and hate speech: An examination of the Nuremberg files case. *Communication Law and Policy*, 10(4), 477–497.

Bartlett, J., & Krasodomski-Jones, A. (2015). C*ounter-speech examining content that challenges extremism online*. Retrieved from https://www.demos.co.uk/wp-content/uploads/2015/10/Counter-speech.pdf

Benesch, S. (2012). *Dangerous speech: A proposal to prevent group violence*. Retrieved from https://worldpolicy.org/wp-content/uploads/2016/01/Dangerous-Speech-Guidelines-Benesch-January-2012

Benesch, S. (2014a). *Countering dangerous speech: New ideas for genocide prevention.* Working paper. Retrieved from https://dangerousspeech.org/countering-dangerous-speech-new-ideas-for-genocide-prevention/

Benesch, S. (2014b). *Flower speech: New responses to hatred online.* Retrieved from https://thenetmonitor.org/research/2014/

Bernstein, M. S., Monroy-Hernández, A., Harry, D., André, P., Panovich, K., & Vargas, G. (2011). 4chan and/b: An analysis of anonymity and ephemerality in a large online community. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, 50–57.

Bitner, M. J., & Obermiller, C. (1985). The elaboration likelihood model: Limitations and extensions in marketing. *ACR North American Advances in Consumer Research, 12,* 420–425.

Boyd, M. S. (2014). (New) participatory framework on YouTube? Commenter interaction in US political speeches. *Journal of Pragmatics*, *72*, 46–58.

Boyle, K. (2001). Hate speech – The United States versus the rest of the world. *Maine Law Review*, *53*, 487–502.

Briñol, P., & Petty, R. E. (2012). A history of attitudes and persuasion research. In A. Kruglanski & W. Stroebe (Eds.), *Handbook of the history of social psychology* (pp. 285–320). New York: Psychology Press.

Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and individual Differences*, *67*, 97–102.

Burnap, P., & Williams, M. L. (2015). Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, *7*(2), 223–242.

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, *119*(2), 197–253.

Cacioppo, J. T., Petty, R. E., & Morris, K. J. (1983). Effects of need for cognition on message evaluation, recall, and persuasion. *Journal of Personality and Social Psychology*, *45*(4), 805–818.

Cammaerts, B. (2009). Radical pluralism and free speech in online public spaces the case of north Belgian extreme right discourses. *International Journal of Cultural Studies*, *12*(6), 555–575.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology*, *39*(5), 752–766.

Chaiken, S. (1987). The heuristic model of persuasion. In M. P. Zanna, J. M. Olson, & C. P. Herman (Eds.), *Social influence: The Ontario Symposium* (Vol. 5, pp. 3–39). Hillsdale, NJ: Erlbaum.

Chandrasekharan, E., Pavalanathan, U., Srinivasan, A., Glynn, A., Eisenstein, J., & Gilbert, E. (2017). You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proceedings of the ACM Human-Computer Interaction*. https://doi.org/10.1145/3134666

Chatzakou, D., Kourtellis, N., Blackburn, J., De Cristofaro, E., Stringhini, G., & Vakali, A. (2017). Measuring #gamergate: A tale of hate, sexism, and bullying. *Proceedings of the 26th international conference on world wide web companion,* 1285–1290.

Chen, C., Zhang, J., Chen, X., Xiang, Y., & Zhou, W. (2015). 6 million spam tweets: A large ground truth for timely Twitter spam detection. *IEEE International Conference on Communications,* 7065–7070.

Cheng, V. T., & Loi, M. K. (2014). Handling negative online customer reviews: The effects of elaboration likelihood model and distributive justice. *Journal of Travel & Tourism Marketing*, *31*(1), 1–15.

Citron, D. K., & Norton, H. L. (2011). Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston University Law Review*, *91*, 1435–1484.

Conte, R., & Paolucci, M. (2014). On agent-based modeling and computational social science. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00668

Costello, M., Hawdon, J., Ratliff, T., & Grantham, T. (2016). Who views online extremism? Individual attributes leading to exposure. *Computers in Human Behavior*, *63*, 311–320.

Crowley, J. P. (2014). Expressive writing to cope with hate speech: Assessing psychobiological stress recovery and forgiveness promotion for lesbian, gay, bisexual, or queer victims of hate speech. *Human Communication Research*, *40*(2), 238–261.

Delgado, R. (1993). Words that wound. A tort action for racial insults, epithets, and name calling. In M. J. Matsuda, C. R. Lawrence III, R. Delgado, K. W. Crenshaw (Eds.), *Words that wound. Critical race theory, assaultive speech, and the First Amendment* (pp. 89-110). New York: Routledge.

Delgado, R., & Stefancic, J. (2014). Hate speech in cyberspace. *Wake Forest Law Review*, *49*, 1–20.

Deuze, M. (1999). Journalism and the web. An analysis of skills and standards in an online environment. *International Communication Gazette, 61*(5), 373–390.

Douglas, K. M., McGarty, C., Bliuc, A.-M., & Lala, G. (2005). Understanding cyberhate: Social competition and social creativity in online white supremacist groups. *Social Science Computer Review*, *23*(1), 68–76.

Douglas, S. C., Kiewitz, C., Martinko, M. J., Harvey, P., Kim, Y., & Chun, J. U. (2008). Cognitions, emotions, and evaluations: An elaboration likelihood model for workplace aggression. *Academy of Management Review*, *33*(2), 425–451.

Elias, S. M., & Loomis, R. J. (2002). Utilizing need for cognition and perceived self-efficacy to predict academic performance. *Journal of Applied Social Psychology*, *32*(8), 1687–1702.

Elin, L., & Davis, S. (2002). *Click on democracy: The internet's power to change political apathy into civic action*. Westview Press, Inc.

Epstein, J. M. (1999). Agent-based computational models and generative social science. *Complexity*, *4*(5), 41–60.

Erjavec, K., & Kovačič, M. P. (2012). "You don't understand, this is a new war!" Analysis of hate speech in news web sites' comments. *Mass Communication and Society*, *15*(6), 899–920.

Festl, R., & Quandt, T. (2013). Social relations and cyberbullying: The influence of individual and structural attributes on victimization and perpetration via the internet. *Human Communication Research*, *39*(1), 101–126.

Foubert, J. D., & Perry, B. C. (2007). Creating lasting attitude and behavior change in fraternity members and male student athletes the qualitative impact of an empathy-based rape prevention program. *Violence Against Women*, *13*(1), 70–86.

Gagliardone, I., Gal, D., Alves, T., & Martinez, G. (2015). *Countering online hate speech*. Paris: UNESCO Publishing.

Gagliardone, I., Pohjonen, M., Beyene, Z., Zerai, A., Aynekulu, G., Bekalu, M., … Teferra, Z. (2016). Mechachal. Online debates and elections in Ethiopia. From hate speech to engagement in social media. http://dx.doi.org/10.2139/ssrn.2831369

Gee, G. (2002). A multilevel analysis of the relationship between institutional and individual racial discrimination and health status. *American Journal of Public Health 92*(4), 615–623.

Gelber, K., & McNamara, L. (2016). Evidencing the harms of hate speech. *Social Identities*, *22*(3), 324–341.

George, C. (2015). Hate speech law and policy. In R. Mansell & P. H. Ang (Eds.), *The International Encyclopedia of Digital Communication and Society* (pp. 305–314). Malden, MA: Wiley Blackwell.

Gerstenfeld, P. B., Grant, D. R., & Chiang, C.-P. (2003). Hate online: A content analysis of extremist internet sites. *Analyses of Social Issues and Public Policy*, *3*(1), 29–44.

Glaser, J., Dixit, J., & Green, D. P. (2002). Studying hate crime with the internet: What makes racists advocate racial violence? *Journal of Social Issues*, *58*(1), 177–193.

Groshek, J., & Cutino, C. (2016). Meaner on mobile: Incivility and impoliteness in communicating contentious politics on sociotechnical networks. *Social Media + Society*, *2*(4), 1–10.

Hanzelka, J., & Schmidt, I. (2017). Dynamics of cyber hate in social media: A comparative analysis of anti-muslim movements in the Czech Republic and Germany. *International Journal of Cyber Criminology*, *11*(1), 143–160.

Haugtvedt, C. P., Petty, R. E., & Cacioppo, J. T. (1992). Need for cognition and advertising: Understanding the role of personality variables in consumer behavior. *Journal of Consumer Psychology*, *1*(3), 239–260.

Heath, B., Hill, R., & Ciarallo, F. (2009). A survey of agent-based modeling practices. *Journal of Artificial Societies and Social Simulation*, *12*(4), 9.

Henry, J. S. (2009). Beyond free speech: Novel approaches to hate on the Internet in the United States. *Information & Communications Technology Law*, *18*(2), 235–251.

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

Hine, G. E., Onaolapo, J., De Cristofaro, E., Kourtellis, N., Leontiadis, I., Samaras, R., ... Blackburn, J. (2017). Kek, Cucks, and God Emperor Trump: A measurement study of 4chan's politically incorrect forum and its effects on the web. *Proceedings of the 11th International AAAI Conference on Web and Social Media*. 92–101.

Jane, E. A. (2014). Your a ugly, whorish, slut. Understanding E-bile. *Feminist Media Studies*, 14(4), 531–546.

Johnson, B. T., & Eagly, A. H. (1989). Effects of involvement on persuasion: A meta-analysis. *Psychological Bulletin*, *106*(2), 290–314.

Johnson, B. T., Maio, G. R., & Smith-McLallen, A. (2005). Communication and attitude change: Causes, processes, and effects. In M. Albarracin & B. T. Johnson (Eds.), *The Handbook of Attitudes. Basic Principles* (pp. 617–669). Mahwah, NJ: Erlbaum.

King, R. D., & Sutton, G. M. (2013). High times for hate crimes: Explaining the temporal clustering of hate-motivated offending. *Criminology*, *51*(4), 871–894.

Kitchen, P. J., Kerr, G., Schultz, D. E., McColl, R., & Pals, H. (2014). The elaboration likelihood model: Review, critique and research agenda. *European Journal of Marketing*, *48*(11/12), 2033–2050.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498.

Laineste, L. (2012). Verbal expressions of aggressiveness in Estonian internet. In L. Laineste, D. Brzozowska, & W. Chłopicki (Eds.) *Estonia and Poland. Creativity and tradition in cultural communication* (pp. 205–220). Tartu: ELM Scholarly Press.

Lange, P. G. (2014). Commenting on YouTube rants: Perceptions of inappropriateness or civic engagement? *Journal of pragmatics, 73*, 53–65.

Lawrence III, C. R. (1990). If he hollers let him go: Regulating racist speech on campus. *Duke Law Journal*, *39*(3), 431–483.

Lee, E., & Leets, L. (2002). Persuasive storytelling by hate groups online examining its effects on adolescents. *American Behavioral Scientist*, *45*(6), 927–957.

Lee, W.-K. (2012). An elaboration likelihood model based longitudinal analysis of attitude change during the process of its acceptance via education program. *Behaviour & Information Technology*, *31*(12), 1161–1171.

Leets, L. (2001). Responses to internet hate sites: Is speech too free in cyberspace? *Communication Law & Policy*, *6*(2), 287–317.

Lodge, M. & Taber, C. (2000). Three steps toward a theory of motivated political reasoning. In A. Lupia, M. D. McCubbins, & S. L. Popkin (Eds.), *Elements of reason, cognition, choice, and the bounds of rationality* (pp. 183-213). Cambridge University Press.

Malthouse, E. C., Calder, B. J., Kim, S. J., & Vandenbosch, M. (2016). Evidence that user-generated content that produces engagement increases purchase behaviours. *Journal of Marketing Management*, *32*(5-6), 427–444.

Maynard, J. L., & Benesch, S. (2016). Dangerous speech and dangerous ideology: An integrated model for monitoring and prevention. *Genocide Studies and Prevention: An International Journal*, *9*(3), 8.

Megarry, J. (2014). Online incivility or sexual harassment? Conceptualising women's experiences in the digital age. In *Women's Studies International Forum, 47*, 46–55.

Mohan, S., Guha, A., Harris, M., Popowich, F., Schuster, A., & Priebe, C. (2017). The impact of toxic language on the health of reddit communities. In M. Mouhoub & P. Langlais (Eds.), *Advances in Artificial Intelligence. AI 2017. Lecture Notes in Computer Science. Volume 10233* (pp. 51-56). Cham: Springer.

Mondal, M., Silva, L. A., & Benevenuto, F. (2017). A measurement study of hate speech in social media. *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. https://doi.org/10.1145/3078714.3078723

Mosler, H.-J., Schwarz, K., Ammann, F., & Gutscher, H. (2001). Computer simulation as a method of further developing a theory: Simulating the elaboration likelihood model. *Personality and Social Psychology Review, 5*(3), 201–215.

Müller, F., van Zoonen, L., & Hirzalla, F. (2014). Anti-Islam propaganda and its effects. *Middle East Journal of Culture and Communication*, *7*(1), 82–100.

Nemes, I. (2002). Regulating hate speech in cyberspace: Issues of desirability and efficacy. *Information & Communications Technology Law*, *11*(3), 193–220.

Orizio, G., Rubinelli, S., Schulz, P. J., Domenighini, S., Bressanelli, M., Caimi, L., & Gelatti, U. (2010). "Save 30% if you buy today". Online pharmacies and the enhancement of peripheral thinking in consumers. *Pharmacoepidemiology and drug safety, 19*(9), 970–976.

Parekh, B. (2006). Hate speech. *Public policy research*, *12*(4), 213–223.

Perry, B., & Olsson, P. (2009). Cyberhate: the globalization of hate. *Information & Communications Technology Law, 18*(2), 185–199.

Petty, R., & Cacioppo, J. (1986). The elaboration likelihood model of persuasion. *Advances in Experimental Social Psychology*, *19*, 124–205.

Petty, R., Cacioppo, J., & Schumann, D. (1983). Central and peripheral routes to advertising effectiveness: The moderating role of involvement. *Journal of Consumer Research*, *10*(2), 135–146.

Petty, R., & Wegener, D. (1998). Attitude change: Multiple roles for persuasion variables. D. Gilbert, S. Fiske, & G. Lindzey (Eds.). *The Handbook of Social Psychology* (pp. 323–390). New York: McGraw-Hill.

Pollock, E. (2009). Researching white supremacists online: methodological concerns of researching hate 'speech'. *Internet Journal of Criminology*, 1–19.

Quandt, T., & Festl, R. (2017). Cyberhate. In P. Rössler (Ed.), *The international encyclopedia of media effects* (pp. 336-344). Malden, Oxford, Chichester: Wiley-Blackwell.

Räsänen, P., Hawdon, J., Holkeri, E., Keipi, T., Näsi, M., & Oksanen, A. (2016). Targets of online hate: Examining determinants of victimization among young Finnish Facebook users. *Violence and Victims*, *31*(4), 708–726.

Richards, R. D., & Calvert, C. (2000). Counterspeech 2000: A new look at the old remedy for bad speech. *Brigham Young University Law Review, 2*, 553–586.

Rösner, L., Winter, S., & Krämer, N. C. (2016). Dangerous minds? Effects of uncivil online comments on aggressive cognitions, emotions, and behavior. *Computers in Human Behavior*, *58*, 461–470.

SanJosé-Cabezudo, R., Gutiérrez-Arranz, A. M., & Gutiérrez-Cillán, J. (2009). The combined influence of central and peripheral routes in the online persuasion process. *CyberPsychology & Behavior*, *12*(3), 299–308.

Schieb, C., & Preuss, M. (2016). Governing hate speech by means of counter-speech on Facebook. *Presentation at the 66th Annual Conference of the International Communication Association*. Fukuoka, Japan.

Schmitt, J. B., Rieger, D., Rutkowski, O., & Ernst, J. (2018). Counter-messages as prevention or promotion of extremism?! The potential role of YouTube recommendation algorithms. *Journal of Communication, 68*(4), 780–808.

Schumann, D. W., Kotowski, M. R., Ahn, H., & Haugtvedt, C. P. (2012). The elaboration likelihood model. In S. Rodgers & E. Thorson (Eds.) *Advertising theory* (pp. 51–68). New York & London: Routledge.

Shane, P. M. (2004). *Democracy online: The prospects for political renewal through the internet*. New York: Routledge.

Shepherd, T., Harvey, A., Jordan, T., Srauy, S., & Miltner, K. (2015). Histories of hating. *Social Media + Society*, *1*(2), 1–10.

Soral, W., Bilewicz, M., & Winiewski, M. (2018). Exposure to hate speech increases prejudice through desensitization. *Aggressive behavior*, *44*(2), 136–146.

Stephenson, M. T., Benoit, W. L., & Tschida, D. A. (2001). Testing the mediating role of cognitive responses in the elaboration likelihood model. *Communication Studies*, *52*(4), 324–337.

Stiff, J. B. (1986). Cognitive processing of persuasive message cues: A meta-analytic review of the effects of supporting information on attitudes. *Communications Monographs, 53*(1), 75–89.

Tsesis, A. (2001). Hate in cyberspace: Regulating hate speech on the internet. *San Diego Law Review, 38*, 817–874.

Voinea, C. F. (2016). *Political attitudes: Computational and simulation modelling*. West-Sussex: John Wiley & Sons.

Vollhardt, J., Coutin, M., Staub, E., Weiss, G., & Deflander, J. (2007). Deconstructing hate speech in the DRC: A psychological media sensitization campaign. *Journal of Hate Studies*, *5*(15), 15–35.

Waldherr, A. (2014). Emergence of news waves: A social simulation approach. *Journal of Com- munication*, *64*(5), 852–873.

Waldron, J. (2012). *The harm in hate speech*. Cambridge & London: Harvard University Press.

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly Journal of Experimental Psychology*, *20*(3), 273–281.

Watts, D. J., & Dodds, P. S. (2007). Influentials, networks, and public opinion formation. *Journal of Consumer Research*, *34*(4), 441–458.

Withers, G. F., Twigg, K., Wertheim, E. H., & Paxton, S. J. (2002). A controlled evaluation of an eating disorders primary prevention videotape using the elaboration likelihood model of persuasion. *Journal of Psychosomatic Research*, *53*(5), 1021–1027.

Withers, G. F., & Wertheim, E. H. (2004). Applying the elaboration likelihood model of persuasion to a videotape-based eating disorders primary prevention program for adolescent girls. *Eating Disorders*, *12*(2), 103–124.

Zerback, T., & Fawzi, N. (2016). Can online exemplars trigger a spiral of silence? Examining the effects of exemplar opinions on perceptions of public opinion and speaking out. *New Media & Society*, *19*(7), 1034–1051.

Ziegele, M., & Jost, P. B. (2016). Not funny? The effects of factual versus sarcastic journalistic responses to uncivil user comments. *Communication Research*. https://doi.org/10.1177/0093650216671854