IASDR 2019
International Association of Societies of Design Research Conference 2019
DESIGN REVOLUTIONS
Manchester School of Art
Manchester Metropolitan University
02-05 September 2019

Design Interventions against Trolling in Social Media: A Classification of Current Strategies Based on Behaviour Change Theories

Kate Sangwon Lee, Huaxin Wei

School of Design, The Hong Kong Polytechnic University, Kowloon, Hong Kong
sangwon.lee@connect.polyu.hk

Trolling in social media has become a serious social issue, and as a result, most social media services have introduced various design interventions to combat it. This study aims to identify effective design interventions against trolling in popular social media services and investigate their approaches based on behaviour change theories in this paper. We have collected 13 representative design intervention cases in six widely popular social media services, and identified their strategies and related design patterns based on existing theories in design for behaviour change. We found two important dimensions—moment of action and required level of user engagement—and classified the 13 design intervention cases based on the two dimensions into four types. The classification of current strategies can serve as an instrument for researchers and designers to analyse and evaluate designing tools against trolling in social media.

Keywords: trolling; social media; socially responsible design; behaviour change; user experience

1 Introduction

Trolling can be defined as “deliberate, deceptive and mischievous attempts to provoke reactions from other users” (Golf-Papez & Veer, 2017). In many contexts of public commenting within social media, trolling is seen as an antisocial behaviour that disrupts discussions within communities. Recently, as the number of social media users has grown rapidly (Statista, 2017b), trolling has become a critical social issue. Recent research (Gammon, 2014) found that trolling frequently happens on chat boards such as Reddit, blog services such as Lifehacker and Jezebel, and popular social media services such as Facebook and Twitter. According to the results of a recent survey in the United States, 38 percent of respondents reported that they saw trolling on social media on a daily basis and 26 percent considered themselves having been victims of trolling (Statista, 2017a).

In response to trolling in social media, designers have created interventions to prevent and constrain it. While some of these interventions and strategies are straightforward by allowing users to use the functions and vary the settings, others take more subtle and hidden approaches in informal and ad hoc ways, such that users might not need to consciously use the features (Shaw, 2018). While there has been much effort and a long history of

Copyright © 2019. Copyright of this paper is the property of the author(s). Permission is granted to reproduce copies of the works for purposes relevant to the IASDR conference, provided that the author(s), source and copyright notice are included on each copy. For other uses, please contact the author(s).

developing tools to combat trolling in social media, little academic research has been conducted to identify the design interventions and strategies in this area.

This study contributes to the field of socially responsible design by examining the current efforts against trolling in social media with the goal of revealing and classifying their strategies. In analysing a corpus of design intervention cases, we attempt to answer two questions: What kinds of design interventions have been implemented for preventing and reacting to trolling in social media and how do these interventions trigger users to combat trolling according to behaviour change theories? Through this study, we arrive at a set of feasible guidelines about how four types of interventions are effective in combating trolling in different contexts of social media services.

2 Backgrounds

2.1 Online trolling

The term "trolling" is defined by Phillips (Phillips, 2014) as "disrupt[ing] a conversation or entire community by posting incendiary statements or stupid questions onto a discussion board...for [the troll's] own amusement, or because he or she was a genuinely quarrelsome, abrasive personality." Hardaker (2010) suggests four distinct elements of trolling: aggression (maliciously annoying others), deception (manipulating online anonymity), disruption (frustrating community members), and success (receiving responses from others).

Hardaker (2010) also differentiates trolling from other online misconduct such as cyberbullying. Cyberbullies often know their victims in real life (Dooley & Cross, 2010), and most of the time, cyberbullying involves a very specific target (Steffgen, König, Pfetsch, & Melzer, 2011). Recently, trolling has become an umbrella term which covers all types of negative online discourse (Golf-Papez & Veer, 2017), including cyberbullying. This study adopts trolling as a specific term which follows Hardaker (2010)'s definition and does not include cyberbullying.

Trolling behaviour happens across diverse subjects, especially sensitive ones such as gender and racism (Mantilla, 2013). The topics that are prone to draw attacks vary from serious ones such as politics and religion to more casual ones such as celebrity and sports (Statista, 2017c). Phillips (2014) suggests that since there is a range of trolling behaviours, from identity-based harassment to political activism, the specific context around each type of trolling should be examined.

Previous studies have discovered that trolling motivations are extremely diverse and depend on context. Sanfilippo et al. (2017) argue that the motivation behind trolling often varies through multiple factors, and the motivation can also vary across the types of trolling. For example, while the motivation behind light-hearted or humorous trolling is mainly to seek enjoyment via disruption and disagreement, more serious trolling is often motivated by ideological or social factors (Sanfilippo, Fichman, & Yang, 2017).

Baccarella et al. (2018) suggest that the main motivation is simple amusement, not pursuit of sharing thought-provoking opinions or discussing about the topics seriously (Baccarella, Wagner, Kietzmann, & McCarthy, 2018). Buckels et al. (2014) discovered that Dark Tetrad traits such as narcissism, Machiavellianism, psychopathy, and sadism positively correlate with the motivation behind trolling (Buckels, Trapnell, & Paulhus, 2014), and Cracker and March (2016) reveal that trolls often regard making social harm as their reward (Craker &

March, 2016). The multidimensionality of trolling implies that the tools against trolling need to be equally diverse.

2.2 Design for behaviour change

The term 'design for behaviour change' emerged in the mid 2000's (Niedderer, 2013). Since then, the power of design as a tool for behaviour change has been discussed by many researchers (Tromp, Hekkert, & Verbeek, 2011) and there have been many types of approaches taken in design for behaviour change.

There are various related terms within design for behaviour change such as nudge, persuasion, and behaviour change theories (Cash, Hartlev, & Durazo, 2017). One of the most developed concepts is persuasive technology. The concept of persuasive technology was introduced in 2003 along with the software-based design for changing behaviour and attitudes through persuasion (Fogg, 2003). In his 2003 paper, Fogg defines persuasive technology as an "interactive technique that is designed to change the attitudes or behaviours or both." Fogg also suggests that current interactive technology can provide many advantages over traditional ones such as broadcasting and print media, because the latter can interact immediately with users' input or intention. Therefore, he argues that designers can use this interactive experience to influence users' motivation and behaviour. Fogg proposes a framework about how technologies can influence users' behaviour through the elements of Tunneling, Tailoring, Suggestion, Self-monitoring, Surveillance, Conditioning, and Reduction. These elements can be foundations in the following studies of design for behaviour change.

2.2.1 Design for responsible behaviour

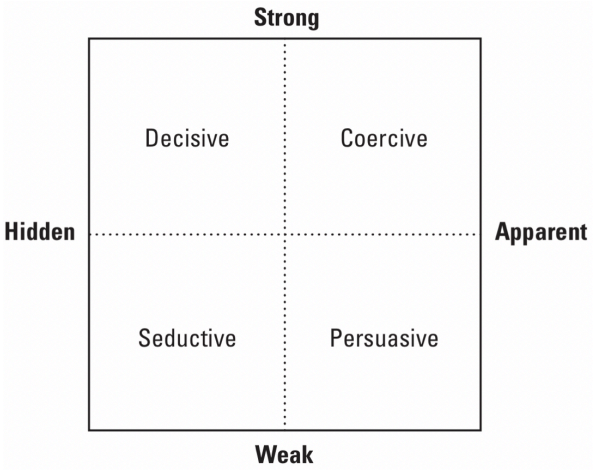
Two application areas of design for responsible behaviour that are related to this research are Design for Sustainable Behaviour (De Medeiros, Da Rocha, & Ribeiro, 2018) and Design against Crime (Press, Erol, Cooper, & Thomas, 2000).

Design for Sustainable Behaviour (DfSB) "aims to reduce products' environmental and social impact by moderating how users interact with them" (Bhamra, Lilley, & Tang, 2011). DfSB is related to both providers and consumers. Providers are encouraged to produce more sustainable products and to reduce the number of environmentally harmful products they offer, while consumers are educated to make environmentally better choices when they purchase goods (Thorpe, 2010). Tang and Tracy (2008) suggest seven strategies to influence consumer behaviour in order to reduce negative social or environmental use: (1) eco-information (2) eco-choice (3) eco-feedback (4) eco-spur (5) eco-steer (6) eco-technical intervention and (7) clever design (Tang & Bhamra, 2008). Since these strategies intend to reduce negative users' behaviour, this study can refer to their scheme.

Design against Crime, introduced by Press et al. (2000), proposes that design against crime aids crime prevention by implementing features to make criminal behaviour less appealing, e.g. by setting up physical barriers against criminal targets or reducing the reward from a crime. This approach discourages a potential criminal's motivation. In the terms of trolling as a type of cybercrime, this current study can partially benefit from this approach.

Based on users' experiences around design for responsible behaviour, Tromp et al. (2011) propose a classification of influence which looks at two important dimensions: salience (apparent or hidden) and force (strong or weak). Based on these two dimensions, they classify four types of influence (see Figure 1): coercive (strong and apparent influence),

persuasive (weak and apparent influence), seductive (weak and hidden influence), and decisive (strong and hidden influence).



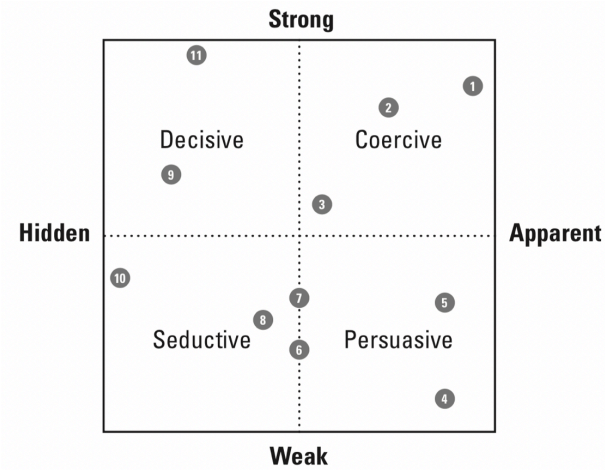


Figure 1. The classification of influence based on intended user experience (left)

Figure 2. Eleven strategies based on the classification of influence (right)

(Source: Tromp et al., 2011)

Based on this classification, they propose 11 design strategies which are effective in influencing responsible user behaviour:

(1) Create a perceivable barrier for undesired behaviour (pain)

(2) Make unacceptable user behaviour overt (shame)

(3) Make the behaviour a prerequisite action to make use of the product function

(4) Provide the user with arguments for the specific behaviour

(5) Suggest actions

(6) Trigger different motivations for the same behaviour

(7) Elicit emotions to trigger action tendencies

(8) Activate physiological processes to induce behaviour

(9) Trigger human tendencies for automatic behavioural responses

(10) Create optimal conditions for specific behaviour

(11) Make the desired behaviour the only possible behaviour to perform.

Figure 2 demonstrates the classification result of the above 11 strategies. We will use these 11 strategies in examining our findings to analyse how they take specific approaches to combat trolling.

2.2.2 "Design with Intent"

Lockton et al. (2010) presented the Design with Intent (DwI) toolkit with 101 design patterns which includes a set of strategies to encourage positive behaviour change. These 101 patterns are grouped under eight lenses: (1) architectural lens, (2) errorproofing lens, (3) interaction lens, (4) ludic lens, (5) perceptual lens, (6) cognitive lens, (7) Machiavellian lens and (8) security lens. In each lens, there are 10 - 15 patterns. This toolkit is an effective

approach to generating initial ideas, conducting brainstorming, and analysing existing products, services, and systems. We will utilize the toolkit to examine the approaches taken by the interventions we have discovered.

3 Method

In this paper, we use 'intervention' as a term to describe a specially designed feature to combat trolling in social media services. We apply the qualitative methods of case study and contents analysis to examine existing design interventions in popular social media services that constrain trolling or help users react to trolling effectively. Our analytical lenses are developed from the 11 behaviour-changing strategies (Tromp et al., 2011) and the Design with Intent toolkit (Lockton, Harrison, & Stanton, 2010). Using these lenses, we investigate the strategy used by each intervention to influence users' experiences and behaviour. Based on the strategies emerged from the investigation, we summarize the results as guidelines for future research and development of socially responsible design.

For our cases, we selected six popular social media services as research objects: Facebook, YouTube, Instagram, Twitter, Tumblr, and Naver. The first five services were selected because of their global popularity (Statista, 2019), while Naver was included because it is the most popular social platform service in South Korea (Statista, 2018) and it has introduced manifold creative and effective interventions against trolling. During the case study, each intervention has been labelled either as proactively preventing trolling (such as muting and blocking) or for restraining trolling (such as deleting and reporting). These two dimensions were discovered from a grounded approach based on the identified interventions through this study.

Our case study included three phases to examine interventions in each service. First, we examined the timeline page (i.e., the main page) and comment sections in each service to discover the general atmosphere and characteristics of each service. Second, we searched several popular news content providers (e.g. the BBC account in Facebook) in each social media service and looked for controversial postings about particularly controversial topics such as racism and gender, as mentioned in section 2.1. We then characterised the kinds of interventions employed to deal with trolling in that comment section. Last, we also studied privacy and security pages in a setting page in each service to look for specially designed interventions against trolling.

4 Results

We have identified the 13 most representative and effective interventions used by six popular social media services against trolling: (1) Real-ID verification, (2) Muting, (3) Disabling comments, (4) Blocking comments from, (5) Comment thread, (6) Thumb up and down, (7) Like, (8) Emotion, (9) Link to profile, (10) Statistics, (11) Report, (12) Automatic filter, and (13) Deleting.

Table 1 presents the 13 design interventions ordered from proactive to reactive and shows which service makes use of which interventions. Comment thread (5) and Report (11) are the most commonly used interventions, being adopted by all six services, while Link to profile (9) and Deleting (13) are the second most used, being adopted by five services. Table 2 summarises the user experiences related to each of the 13 interventions, and analyses

them using Tromp et al.'s 11 behaviour-changing strategies and the 101 patterns in the Dwl toolkit of Lockton et al.

Table 1. Social media services and their interventions against trolling. An "o" indicates that the service makes use of the intervention.

		Facebook	YouTube	Instagram	Twitter	Tumblr	Naver
Proactive	1 Real-ID verification						o
	2 Muting			o	o		
	3 Disabling comments		o				
	4 Blocking comments from			o			
	5 Comment thread	o	o	o	o	o	o
	6 Thumb up and down		o				o
	7 Like			o	o		
	8 Emotion	o					
	9 Link to profile	o	o	o	o	o	
Reactive	10 Statistics						o
	11 Report	o	o	o	o	o	o
	12 Automatic filter				o		
	13 Deleting	o	o	o	o		o

Table 2. 13 interventions against trolling and user experiences

1 Real-ID verification

Adopted by Naver

How it works

- Users verify their real names with id numbers in order to join the designated online community.

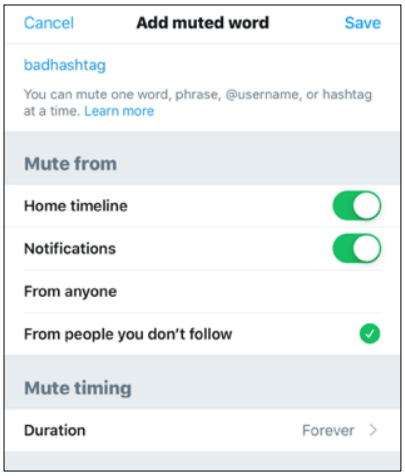
Strategies/Classifications adopted from Tromp et al.

- (3) Make the behaviour a necessary activity to perform to make use of the product function (Coercive): Without verification, users are not allowed to join the online community.

Lenses/Patterns adopted from the Dwl toolkit

- Errorproofing lens (Defaults)
- Cognitive lens (Do as you're told)
- Security lens (Who or what you are, Surveillance)

2 Muting



Adopted by Instagram, Twitter

How it works

- Users can register specific hate words to be muted; comments containing the muted words are not shown.

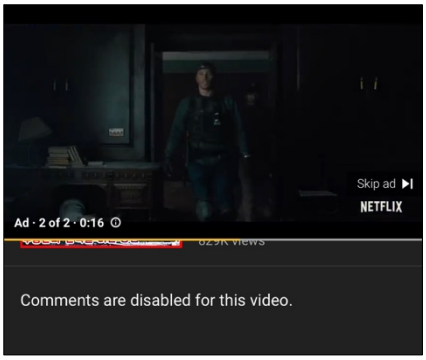
Strategies/Classifications adopted from Tromp et al.

- (5) Suggest actions (Persuasive): Suggest to users the way to avoid trolling.

Lenses/Patterns adopted from the Dwl toolkit

- Interaction lens (Tailoring)

3 Disabling comments



Adopted by YouTube, Instagram

How it works

- Contents providers who do not want to allow any negative comments can disable all comments.

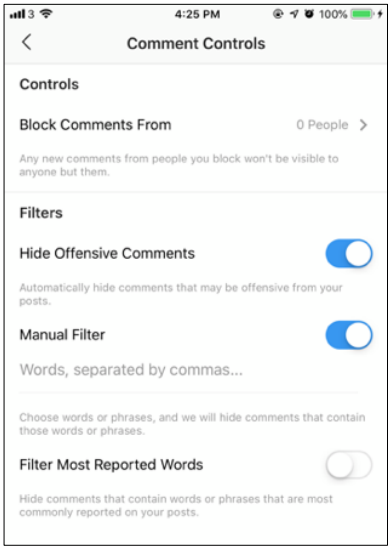
Strategies/Classifications adopted from Tromp et al.

- (1) Create a perceivable barrier for undesired behaviour (Coercive): Disable possible undesired behaviour.

Lenses/Patterns adopted from the Dwl toolkit

- Architectural lens (Feature deletion, Hiding things)
- Security lens (Surveillance)

4 Blocking comments from



Adopted by Instagram

How it works

- Users who want to avoid comments from a specific source can specify the source id and block all comments from that source.

Strategies/Classifications adopted from Tromp et al.

- (5) Suggest actions (Persuasive): Suggest to users a way to avoid a troll.

Lenses/Patterns adopted from the Dwl toolkit

- Interaction lens (Tailoring)
- Security lens (Peerveillance)

5 Comment thread

Adopted by Facebook, YouTube, Instagram, Twitter, Tumblr, Naver

How it works

- Users can reply to comments they want to discuss, sharing their opinions as to whether the comments are from trolls.

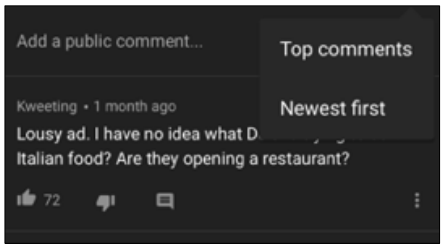
Strategies/Classifications adopted from Tromp et al.

- (4) Provide the user with arguments for specific behaviour (Persuasive): Users can argue with trolls using this feature.
- (6) Trigger different motivations for the same behavior (Persuasive & Seductive): By expressing their own ideas, users can discuss controversial topics, and find compromising points.

Lenses/Patterns adopted from the Dwl toolkit

- Interaction lens (Peer feedback)
- Ludic lens (Playfulness): Sometimes users interact with each other, and they can create unexpected results.
- Security lens (Peerveillance)

6 Thumb up and down



Adopted by YouTube, Naver

How it works

- Users can evaluate each comment with this feature. By looking at the evaluations as they accumulate, users can possibly tell if a comment is trolling.

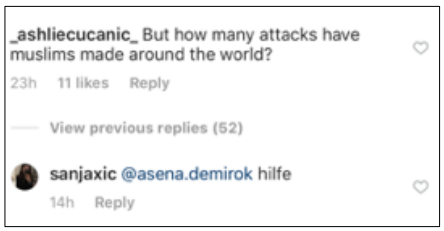
Strategies/Classifications adopted from Tromp et al.

- (6) Trigger different motivations for the same behavior (Persuasive & Seductive): by deploying their own ideas, users can discuss controversial topics, and find possible compromising points.
- (7) Elicit emotions to trigger action tendencies (Persuasive & Seductive): users can express their feelings about comments.
- (9) Trigger human tendencies for automatic behavioural responses (Decisive): People enjoys rewards and praises by getting thumb ups.

Lenses/Patterns adopted from the Dwl toolkit

- Interaction lens (Peer feedback)
- Ludic lens (Playfulness): Users can interact with each other, creating unexpected results.
- Security lens (Peerveillance)

7 Like



Adopted by Instagram, Twitter, Tumblr

How it works

- This is a simplified version of thumb up and down. With it, users can tell which comments are most popular. However, it does not communicate negative reactions.

Strategies/Classifications adopted from Tromp et al.

- (4) Provide the user with arguments for specific behaviour (Persuasive): users can argue with this feature.
- (7) Elicit emotions to trigger action tendencies (Persuasive & Seductive)

Lenses/Patterns adopted from the Dwl toolkit

- Interaction lens (Peer feedback)
- Ludic lens (Playfulness, Levels, Rewards, Scores) Security lens (Peerveillance)

8 Emotion

Adopted by Facebook

How it works

- Users can express diverse emotions with this feature: like, love, haha, yay, wow, sad, and angry. By looking at the three emotions most chosen, especially if they are angry emotions, users can detect trolling.

Strategies/Classifications adopted from Tromp et al.

- (2) Make unacceptable user behaviour overt (shame) (Coercive)
- (7) Elicit emotions to trigger action tendencies (Persuasive & Seductive)

Lenses/Patterns adopted from the Dwl toolkit

- Interaction lens (Peer feedback)
- Ludic lens (Playfulness, Levels, Rewards, Scores) Security lens (Peerveillance)

9 Link to profile

Adopted by Facebook, YouTube, Instagram, Twitter, Tumblr

How it works

- These five social media services provide the commenters' profile picture, and a link to their profiles as well. So, commenters may feel more responsible than if they were commenting anonymously.

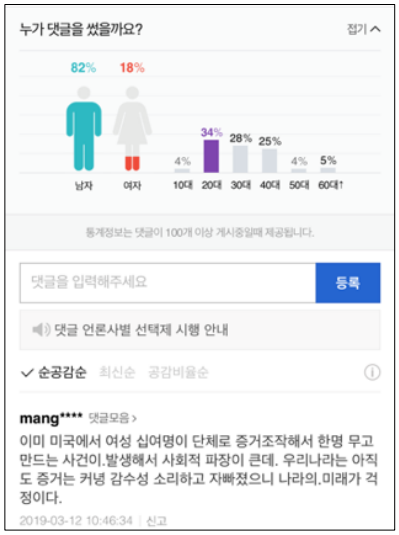
Strategies/Classifications adopted from Tromp et al.

- (2) Make unacceptable user behaviour overt (shame) (Coercive)
- (3) Make the behaviour a necessary activity to perform to make use of the product function (Coercive)
- (11) Make the desired behaviour the only possible behaviour to perform (Decisive)

Lenses/Patterns adopted from the Dwl toolkit

- Perceptual lens (Transparency, Watermarking)
- Security lens (Peerveillance, Threat to property)

10 Statistics



Adopted by Naver

How it works

- Demographic statistics are shown to reflect all comments. These are especially effective in detecting gender-trolling by providing the gender ratio of the commenters. The male-to-female ratio is usually extremely high after an article that attracts many gender-trolling comments.

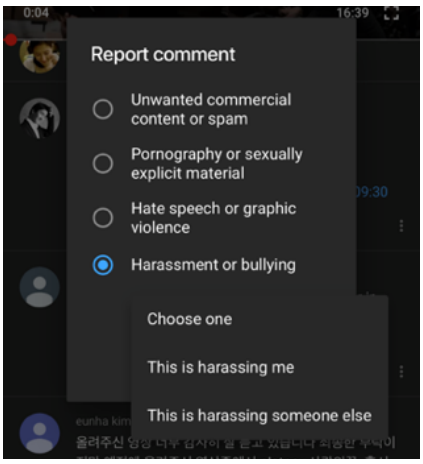
Strategies/Classifications adopted from Tromp et al.

- (4) Provide the user with arguments for specific behaviour (Persuasive): the user is provided with objective information about the demographic of previous commenters.

Lenses/Patterns adopted from the Dwl toolkit

- Interaction lens (Real-time feedback, Summary feedback)
- Perceptual lens (Transparency)
- Security lens (Peerveillance)

11 Report



Adopted by Facebook, YouTube, Instagram, Twitter, Tumblr, Naver

How it works

- Users can report a trolling comment by clicking a link, and they can provide their reasons.

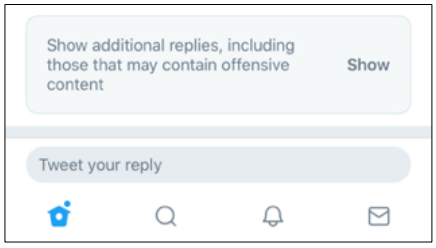
Strategies/Classifications adopted from Tromp et al.

- (4) Provide the user with arguments for specific behaviour (Persuasive)
- (5) Suggest actions (Persuasive): the report feature can be a trigger for users to behave more consciously and carefully, since otherwise, the comment might be reported.

Lenses/Patterns adopted from the Dwl toolkit

Security lens (Peerveillance)

12 Automatic filter



Adopted by Twitter

How it works

- When the service detects trolling, it hides the comments; the comments are shown only if a user asks to see them.

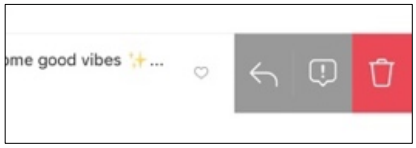
Strategies/Classifications adopted from Tromp et al.

- (5) Suggest actions (Persuasive): when something undesirable happens, suggest an option not to face it.

Lenses/Patterns adopted from the Dwl toolkit

Architectural lens (Hiding things)

13 Deleting



Adopted by Facebook, YouTube, Instagram, Twitter, Naver

How it works

- When the contents provider detects trolling, it can delete the comments by its authority.

Strategies/Classifications adopted from Tromp et al.

- (10) Create optimal conditions for specific behavior (Seductive): force the comment area trolling-free by service provider

Lenses/Patterns adopted from the Dwl toolkit

Security lens (Surveillance)

Figure 3 classifies the 13 interventions along two important dimensions: ‘Moment of action’ and ‘Required level of user engagement’. ‘Moment of action’ describes at which moment the

intervention is used for trolling management. 'Required level of user engagement' describes the degree of user engagement required for each intervention, ranging from active engagement (e.g. comment thread, report) to a more passive attitude (e.g. statistics, link to profile). Figure 3 then identifies four types of design interventions against trolling in social media: *Preventive, Precautionary, Participatory,* and *Administrative*, which can serve as guidelines for each of the four types of design interventions for future designers. In the following, we describe the four types in relation to their intervention types that are numbered in Table 2 and Figure 3.

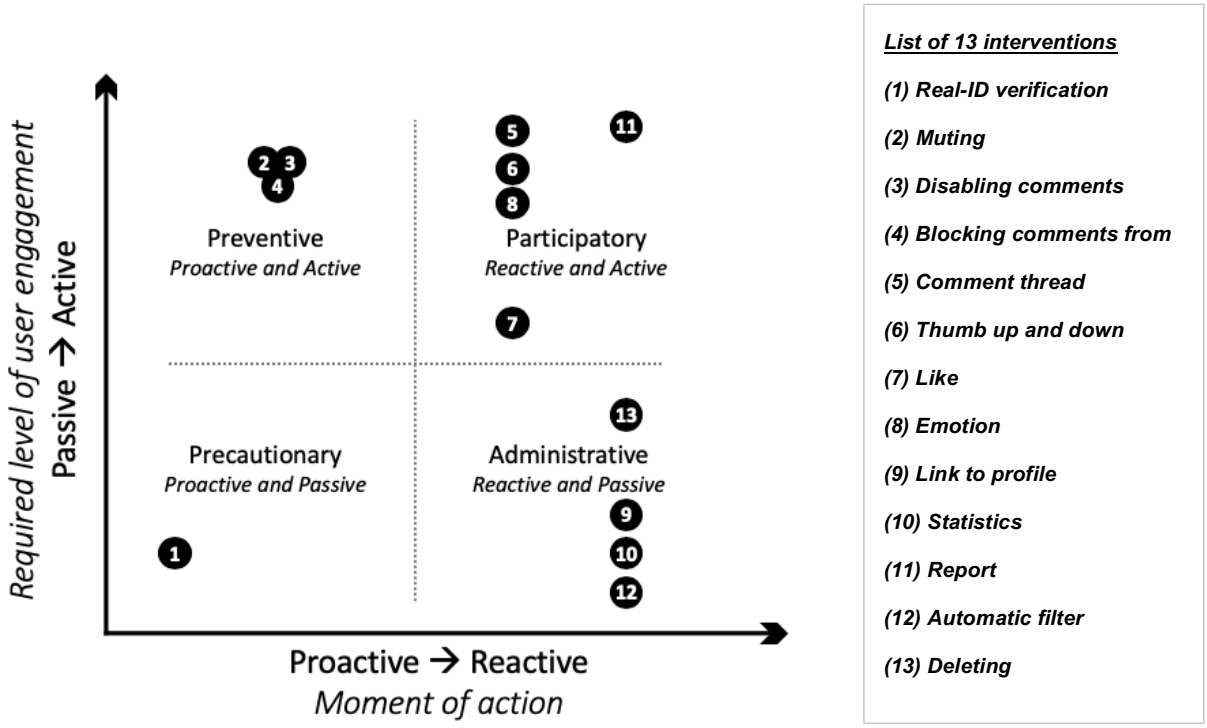


Figure 3. Four types of design interventions against trolling in social media

Preventive interventions (proactive and active): *Muting (2), Disabling comments (3), Blocking comments from (4)*. These interventions help users avoid trolling with predictable patterns or hate wordings. If some visible pattern of trolling is identified, then a tool to prevent it should be provided. If the contents provider wants to avoid all possible kinds of trolling, then a comment disabling option should be provided. These types of interventions are necessary for all kinds of social media, because they help users use the service in a safe and secure way.

Precautionary interventions (proactive and passive): *Real-ID verification (1).*
This type of intervention functions proactively to track down and control commenters, making users feel that they are under the control of the service provider (surveillance). However, service providers should be careful in adopting this intervention and avoid designing complicated steps so that it does not interfere with users' access in the early stages of service usage.

Participatory interventions (reactive and active): *Comment thread (5), Thumb up and down (6), Like (7), Emotion (8), Report (11).* Since social media is a platform that congregates many users, it can utilise the users' active participation and collective intelligence (peerveillance) to deal with trolling. Social media can provide tools for users to express their feelings or emotions in a simple and easy way (e.g. positive or negative). Once expressed, it is important that the feelings and emotions are presented effectively to other users. This strategy will make trolling visible to users so that they themselves can take the authority to constrain trolling. We recommend this type of intervention for services where user participation actively occurs.

Administrative interventions (Reactive and Passive): *Link to profile (9), Statistics (10), Automatic filter (12), Deleting (13).*
Service administrators can decide to enter the comments area and provide relevant information or surveillance to make the area clean from trolling. Service administrators should watch what is happening there and react in proper ways. They can adopt an Artificial Intelligence filter to hide trolling automatically or delete trolling comments manually. They can provide informative features to help users recognize trolling easily. This strategy might require intense administrative time and effort, and therefore it is most appropriate for those social media services where severe trolling often happens.

5 Discussion

We expect designers of social media services to use our proposed classification and strategies in a way that depends on the particular stage of development of the social media.

In the initial stage of service development, designers can refer to our 13 representative interventions against trolling for ideation or design decision making. They will learn how these interventions might work in their service and affect their users' experience. Our findings can also be used as the basis for brainstorming sessions leading to new interventions against trolling.

If a social media service is already in its mature stage, designers can utilize our strategies as a tool to examine their existing anti-trolling interventions with regards to the contexts and characteristics of their service. For example, if the users of a service have a tendency to participate actively in comment sections, designers can check whether they embody *participatory* interventions in their service by empowering their users in an effective way.

This study has the limitation that certain social media applications might not be covered by our framework of strategies. Crucially, our findings mainly focus on users who want to avoid and constrain trolling. Future study can be carried out to reveal possible interventions to moderate both the motivations and behaviour of online trolls themselves.

With our findings, this paper can be a foundation for future work. To broaden the study, possible future directions include: (1) adding an empirical user study to examine the effectiveness of proposed strategies, and (2) including other types of social media service platforms besides mobile applications, such as mobile web or PC web-based ones (e.g., GeoExpat). Our classification and findings should stay fluid and be updated over the time with new design phenomena emerging.

6 Conclusion

This paper presents the analysis of 13 design interventions against trolling in social media and describes the user experiences and the interventions' implications based on existing design strategies for behavior change. As a result, we propose four intervention types—*preventative, precautionary, participatory,* and *administrative*. We hope the classification can not only serve as an evaluative tool for design researchers but also inform social media designers who want their designs to be socially responsible.

7 References

Baccarella, C. V., Wagner, T. F., Kietzmann, J. H., & McCarthy, I. P. (2018). Social media? It's serious! Understanding the dark side of social media. European Management Journal, 36(4), 431-438.

Bhamra, T., Lilley, D., & Tang, T. (2011). Design for sustainable behaviour: Using products to change consumer behaviour. The Design Journal, 14(4), 427-445.

Buckels, E. E., Trapnell, P. D., & Paulhus, D. L. (2014). Trolls just want to have fun. Personality and Individual Differences, 67, 97-102. doi:10.1016/j.paid.2014.01.016

Cash, P. J., Hartlev, C. G., & Durazo, C. B. (2017). Behavioural design: A process for integrating behaviour change and design. Design Studies, 48, 96-128.

Craker, N., & March, E. (2016). The dark side of Facebook®: The Dark Tetrad, negative social potency, and trolling behaviours. Personality and Individual Differences, 102, 79-84.

De Medeiros, J. F., Da Rocha, C. G., & Ribeiro, J. L. D. (2018). Design for sustainable behavior (DfSB): Analysis of existing frameworks of behavior change strategies, experts' assessment and proposal for a decision support diagram. Journal of Cleaner Production, 188, 402-415. doi:https://doi.org/10.1016/j.jclepro.2018.03.272

Dooley, J. J., & Cross, D. (2010). Cyberbullying versus face-to-face bullying: A review of the similarities and differences. Journal of Psychology, 217, 182-188.

Fogg, B. (2003). How to Motivate & Persuade Users. CHI 2003: New Horizons.

Gammon, J. (2014). Over a quarter of Americans have made malicious online comments.

Golf-Papez, M., & Veer, E. J. J. o. M. M. (2017). Don't feed the trolling: rethinking how online trolling is being defined and combated. 33(15-16), 1336-1354.

Hardaker, C. (2010). Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. In: Walter de Gruyter GmbH & Co. KG.

Lockton, D., Harrison, D., & Stanton, N. A. (2010). The Design with Intent Method: A design tool for influencing user behaviour. Applied ergonomics, 41(3), 382-392.

Niedderer, K. (2013). Mindful design as a driver for social behaviour change. Paper presented at the Proceedings of the IASDR Conference 2013.

Phillips, W. (2014). To Fight Trolls, Focus on Actions and Context. Retrieved from https://www.nytimes.com/roomfordebate/2014/08/19/the-war-against-online-trolls/to-fight-trolls-focus-on-actions-and-context

Press, M., Erol, R., Cooper, R., & Thomas, M. (2000). Design against crime: defining new design knowledge requirements.

Sanfilippo, M. R., Fichman, P., & Yang, S. (2017). Multidimensionality of online trolling behaviors. The Information Society, 34(1), 27-39. doi:10.1080/01972243.2017.1391911

Shaw, F. (2018). Beyond 'report, block, ignore': Informal responses to trolling and harassment on social media. In The Routledge Companion to Media and Activism (pp. 395-403): Routledge.

Statista. (2017a). How often do you see internet trolling on the following types of media?: Witnessing internet trolling on selected media in the U.S. 2017.

Statista. (2017b). Number of social network users worldwide from 2010 to 2021 (in billions).

Statista. (2018). Most popular online properties in South Korea as of May 2018, by number of unique visitors (in millions).

Statista. (2019). Most famous social network sites worldwide as of January 2019, ranked by number of active users (in millions).

Steffgen, G., König, A., Pfetsch, J., & Melzer, A. (2011). Are cyberbullies less empathic? Adolescents' cyberbullying behavior and empathic responsiveness. Cyberpsychology, Behavior, and Social Networking, 14(11), 643-648.

Tang, T., & Bhamra, T. (2008). Changing energy consumption behaviour through sustainable product design. Paper presented at the DS 48: Proceedings DESIGN 2008, the 10th International Design Conference, Dubrovnik, Croatia.
Thorpe, A. (2010). Design's role in sustainable consumption. Design Issues, 26(2), 3-16.
Tromp, N., Hekkert, P., & Verbeek, P.-P. (2011). Design for socially responsible behavior: a classification of influence based on intended user experience. Design Issues, 27(3), 3-19.

About the Authors:

Kate Sangwon Lee: a PhD student in the School of Design at The Hong Kong Polytechnic University. She worked as a User Experience designer and researcher since 2008. Kate holds master's degree in Culture Technology at KAIST, Bachelor's degree at Seoul National University.

Huaxin Wei: an associate professor in the School of Design at The Hong Kong Polytechnic University, Huaxin's research interests include interaction design, game design analysis, and interactive narrative. Huaxin holds a PhD in Interactive Arts and Technology at Simon Fraser University, Canada.

Acknowledgement: The authors would like to thank Dr. Leslie Schwartz for giving helpful opinions.