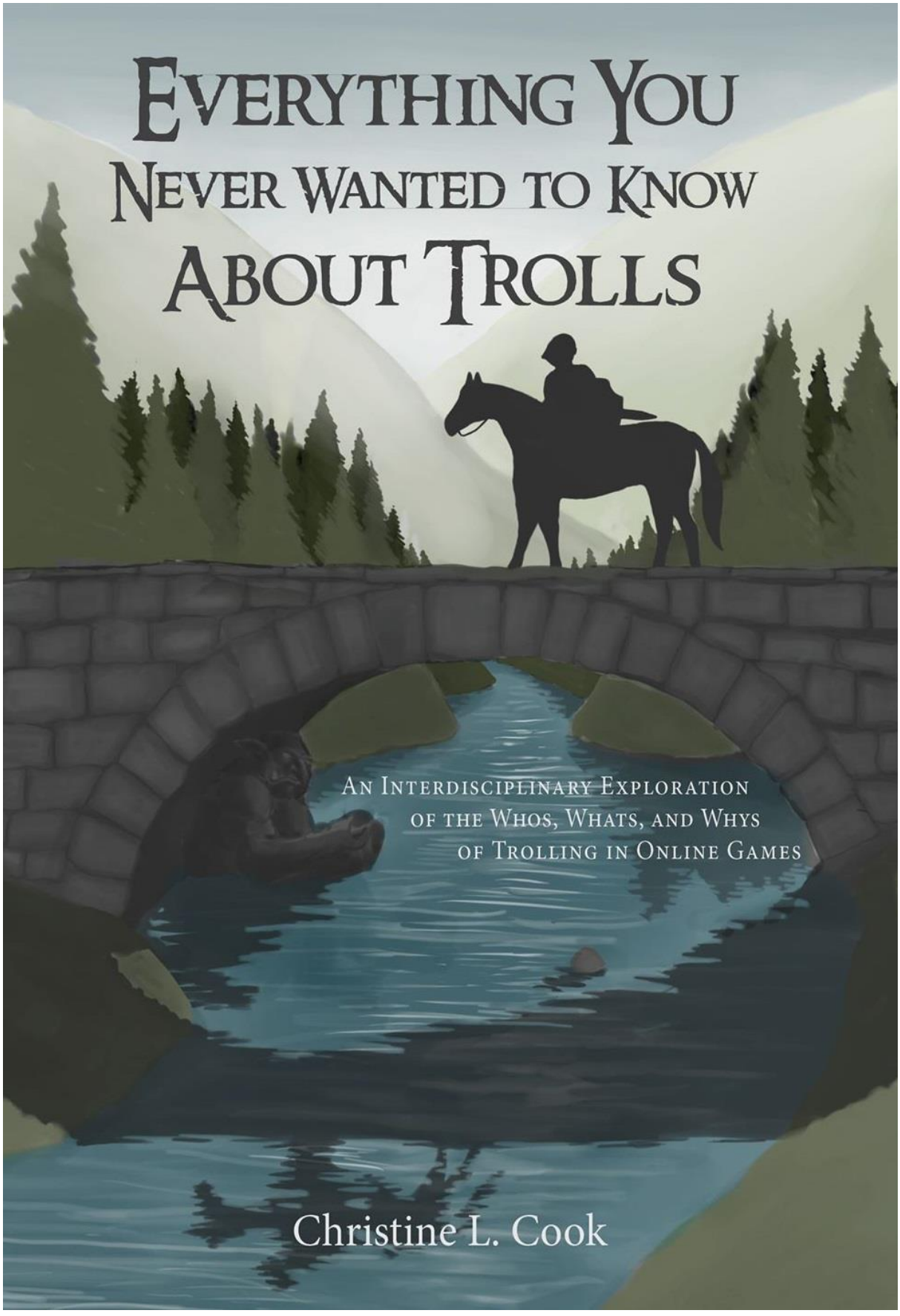
Everything You Never Wanted to Know about Trolls

Cook, Chrissy

Publication date:
2021

Document Version
Publisher's PDF, also known as Version of record


Everything You Never Wanted to Know About Trolls
An Interdisciplinary Exploration of the Whos, Whats, and Whys of Trolling in Online Games
Christine L. Cook

Everything You Never Wanted to Know about Trolls

An Interdisciplinary Exploration of the Who's, What's, and Why's of Trolling in Online Games

A doctoral dissertation by
Christine "Chrissy" Cook

Cover art by
Karl Gruenewald

Everything You Never Wanted to Know about Trolls

An Interdisciplinary Exploration of the Who's, What's, and Why's of Trolling in Online Games

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan Tilburg University
op gezag van de rector magnificus, prof. dr. W.B.H.J. van de Donk, in het
openbaar te verdedigen ten overstaan van een door het college voor
promoties aangewezen commissie in de Portrettenzaal van de Universiteit op
vrijdag 22 januari 2021 om 13.30 uur

door

Christine Linda Cook,

geboren op 20 january, 1992 te Moncton, Canada

Promotores:	prof. dr. J. Schaafsma (Tilburg University) prof. dr. M.L. Antheunis (Tilburg University)
leden promotiecommissie:	prof. dr. N.C. Krämer (University Duisburg-Essen) dr. B.H. Detenber (Nanyang Technological University Singapore) prof. dr. H. Vandenbosch (University of Antwerp) prof. dr. T. Hartmann (Vrije Universiteit Amsterdam) prof. dr. M.G.J. Swerts (Tilburg University)

ISBN: 978-1-7774987-0-2

Naam drukkerij (Printer): KKP Canada (www.kkpcanada.ca)

Table of Contents

Chapter 1: Introduction

To say that gaming as a pastime has become ubiquitous would be an understatement. In 2018, the sale of video games in the United States reached a record-breaking $43.4 billion USD (Entertainment Software Association, 2019), not including the equally strong markets in South Korea (Sue & Mintegral, 2019), China (Blazyte, 2019), and Japan (Newzoo, 2018). Even in countries like the Netherlands with much smaller populations than the USA or China, gaming as a hobby and as a profession is on the rise, with developers and gamers alike achieving global success and raking in significant winnings (Dutch Games Association, 2019). People around the world are gaming, and doing so regularly. However, this global ubiquity is not without its downsides. Though the enjoyment of video games is spreading throughout the world, so is one of gaming's darkest trends: trolling.

Trolling has been discussed and defined in a variety of ways by both the media and by academics. It has been particularly vilified by journalists like Joel Stein (2016), who when writing for Time magazine called trolls "the culture of hate". Scholar Whitney Phillips (2016) echoed this idea when she titled her book about trolls and trolling "This is why we can't have nice things", emphasizing trolling culture as attacking and even consuming the mainstream. That said, though media and academia generally agree that trolling is undesirable, there is considerably less consensus when it comes to what trolling actually *is*. Cocomello (2016) when writing for online magazine MyGaming describes trolling in games as "an art" in which the troll is gratified at the expense of their victim. Donath (1999), in one of the earliest scientific articles on the topic, describes trolling as "a game about identity deception" (p. 6). However, Fichman and Sanfilippo (2014) give trolling a much darker tone, defining it as "deviant and antisocial online behavior in which the deviant user acts provocatively and outside of normative expectations within a particular community" (p. 163). Types of trolling in online games have been described in literature as ranging from loudly (and badly) singing

into a microphone to irritate teammates, to viciously insulting the players around them as people and as gamers until they leave the game (Thacker & Griffiths, 2012). We as researchers have plenty of ideas about what trolling *could* be, but what *is* it really? Beyond the question of what constitutes trolling, whom does it affect, and how do these effects change depending on how trolling occurs and who is involved?

It is this first question of "what is trolling" that sparked the initial idea for this dissertation: to address the most basic, fundamental questions of trolling from a scientific, multi-method perspective. Although when I undertook the dissertation there were a multitude of ideas about trolling being presented in both the media and academia, these were all discrete; one or two articles would appear in one discipline, then one or two in another, but there was no one systematic examination of trolling as a complete phenomenon. Each discipline was tackling its own individual questions of interest related to trolling, which creates a body of knowledge when gathered together, but leaves the foundational questions of trolling unanswered. The goal of the present work is to rectify this gap by exploring the various perspectives involved in trolling – that of the troll, the victim(s), and the bystander(s) – to understand a) what constitutes trolling, b) why trolls do what they do, and c) how victims and bystanders (the surrounding community) react to various types of trolling in various contexts. Together, the answers to these questions should help us better understand what contributes to the spread of trolling, primarily in online games, and what inhibits this spread. I aim to build on existing knowledge by taking into account all the descriptors of trolls and trolling listed in the earliest works on the subject (e.g., Herring, Job-Sluder, Scheckler, & Barab, 2002) and the most recent (e.g., Graham, 2019) across the myriad disciplines involved in trolling research. Then, by looking at actual trolls, victims, and bystanders over the course of my four studies, we can see which of these variables emerge as critically important in real and simulated trolling situations. By answering these questions, I aim to build a bridge

between the various findings from across academia and provide in turn sets of variables that have a) been shown to have significance for one of the three actors in trolling situations, and b) been validated for the online gaming context.

Trolls and their Behaviour

The wealth of descriptive studies available in trolling, are almost universally focused on the troll as a person. Some examples include Suler and Phillips' (1998) piece, which lists the types of what we would today call trolls (e.g., deviant enclave members, sleepers) and the kinds of trolling they can do in chat communities (e.g., abusive blocking, graffiti), or Herring and colleagues' (2002) list of the various techniques a single troll employed (e.g., repetition, excessive use of questions) to wreak havoc in an online feminist forum. These studies provide very relevant, critical observational data about trolls and their behaviour, and serve as the basis for later work looking to take a deeper look at trolls' motivations. One such project is Phillips' (2011) infiltration of a community of trolls. Her work includes interviews with self-confessed trolls in which they talk about motivations to troll, such as personal satisfaction and raising awareness (Phillips, 2011; 2013; 2016). However, despite the fact that trolls and trolling behaviour are the main focus of the bulk of trolling literature (e.g., Buckels et al., 2014; Hardaker, 2010; Shachaf & Hara, 2010; Thacker & Griffiths, 2012), there is still considerable debate concerning which behaviours specifically constitute trolling, and which do not.

One particularly powerful illustration of this question can be found by looking at the various definitions given to trolling. One of the most-cited papers about trolls and trolling, Buckels and colleagues' (2014) "Trolls just want to have fun" article, defines it as follows: "Online trolling is the practice of behaving in a deceptive, destructive, or disruptive manner in a social setting on the Internet with no apparent purpose" (p. 97). There are a few key

concepts in this definition, such as the idea of destroying or ruining something, and the idea of wanton-ness, that this destruction is essentially random. This contrasts majorly with how trolling is perceived by Thacker and Griffiths (2012): "Trolling is an act of intentionally provoking and/or antagonizing users in an online environment that creates an often desirable, sometimes predictable, outcome for the troll" (p. 18). According to these researchers, trolling is fundamentally instrumental in nature, the opposite of wanton destruction. There is also no specific mention of deception or destruction; Thacker and Griffiths (2012) state only that what they do, at the least, provokes a victim. Even with our lists of possible trolling behaviours (Suler & Phillips, 1998; Thacker & Griffiths, 2012), literature is not clear on what *makes those behaviours trolling*. We do not yet know what makes an internet flame (see O'Sullivan & Flanagin, 2003) a case of trolling, for instance; we only know that flaming is sometimes called trolling when it happens in online games (see Thacker & Griffiths, 2012). To find out what really makes a case of trolling a case of trolling, I turn to asking the trolls themselves directly in Chapter 2 and looking at actual cases of trolling that have been reported by victims and bystanders in Chapter 3. In this way, we can pinpoint key variables that characterize trolling interactions and determine the critical elements that need to be present in order for behaviour to constitute trolling.

Trolls and their Motivations

Another subject that still requires investigation is the motivations of trolls. We can see this ongoing question when we look at trolling definitions again, for example Thacker and Griffiths' (2012) emphasizing the idea of obtaining pleasure, and Buckels and colleagues' (2014) emphasizing the desire to destroy. This debate is also highlighted in the question of trolling behaviour's origins. One recent article posits that trolling is a form of boundary maintenance (Graham, 2019). According to this position, trolls serve as the guardians of their respective communities, testing potential members and outsiders seeking entry. Others should

have to prove themselves worthy before they are admitted, and the trolls are those who test. Sometimes this can be an offensive tactic as well, with trolls asserting the dominance of their community by infiltrating or attacking other communities perceived as being weaker or less important (Graham, 2019). Another postulation is that behaviours like trolling are intrinsic to the internet's infrastructure (Kerr & Lee, 2019). By the simple nature of the internet and its affordances and characteristics (e.g., anonymity), trolling happens organically and will not stop unless we change the fundamental structure of the internet. Unlike the boundary maintenance position, this one has no implicit motivation for trolls. Instead, it would suggest that the desire to troll is intrinsic, and the collection of affordances that make up the internet simply allows it to be expressed (Kerr & Lee, 2019). Contrasting these positions, the popular conceptualization of trolling presented commonly in the media is that trolling is simply the evolution of practical jokes and pranking behaviour (Bergtau, 2014), implying that trolls are hedonistically motivated.

These different propositions reflect the kind of information that we have about trolls' motivations: inferred or deduced. This is because most of it comes from case studies, two of the most well-known being Phillips' work (2016) and Herring and colleagues' (2002) study of a troll in an online feminist forum. As previously mentioned, Phillips' (2011; 2013; 2016) focus was on trolls as shapers of digital culture, and many of her publications compare elements of trolling culture with that of mainstream culture. Herring and colleagues' (2002) study also focuses on a troll operating on an online forum, although it was a feminist forum instead of a political discussion space. When we look at trolling interactions from an outsider's perspective, we can only make educated guesses when it comes to trolls' motivations. This is one of the major reasons that I conducted a study in which trolls are questioned directly: to determine exactly what motivates them to do what they do (Chapter

2). By conducting in-depth interviews with trolls, we can find out which motivations deduced from literature reflect the actual thoughts and desires of real-life trolls.

Victims and Bystanders of Trolling

The last topic the present dissertation addresses is that of the different perspectives involved in trolling research. As previously explained, a major portion of trolling research focuses on either the person of the troll or the act of trolling (e.g., Buckels et al., 2014; Hardaker, 2010; Shachaf & Hara, 2010; Thacker & Griffiths, 2012). However, trolling is fundamentally social, and requires at least a victim, if not onlookers (Buckels et al., 2014; Fichman & Sanfilippo, 2014; Thacker & Griffiths, 2012). These other perspectives – that of the victim(s) and bystander(s) – would be a valuable addition to trolling scholarship, as these are the people who could potentially either stop the trolling, say by reporting the individual to authorities (see Fox, Gilbert, & Tang, 2018), or encourage further trolling (see Thacker & Griffiths, 2012).

Some descriptive studies of trolling like Herring and colleagues' (2002) case study do document the actions of people other than the troll, but the emphasis of the study is still on the actions of the troll. Other studies using a more experimental methodology often involve people who could potentially be victims or bystanders of trolling incidents (e.g., McCosker, 2014), but participants were not screened for that information, so there is no way to tell who is and is not, and how that status may affect their reactions to the experimental manipulations (e.g., Maltby et al., 2015). Because of this, a lot of the information that we have about victims and bystanders comes from other fields, primarily cyberbullying (e.g., Blackburn & Kwak, 2014; Chesney, Coyne, Logan, & Madden, 2009; Kwak, Blackburn, & Han, 2015). Practically, this means that we know very little about how bystanders and victims of trolling actually feel and think, despite their potential importance in trolling interactions (e.g.,

Johnson, Cooper, & Chin, 2009). This dissertation therefore aims to expand the perspectives included in trolling literature, not only by actively testing bystanders to and victims of trolling (Chapters 4 and 5), but also by exploring how existing theories of computer-mediated communication and psychology apply to the actors in trolling interactions.

In fact, existing research suggests that bystanders and victims can add as much variety to a trolling situation as the troll can by varying their method of trolling. The way victims or bystanders react to trolling can vary wildly, even if the type of trolling remains consistent (see Herring et al., 2002, and Thacker & Griffiths, 2012 for examples). As the internet is a sort of global playground, one source of variety among victims and bystanders may be their offline cultural context. Researchers interested in cross-cultural psychology, for instance, have repeatedly found that a person's cultural context influences their reactions to directed aggression (Cohen, 1998; Hashimoto & Yamagishi, 2013; Park et al., 2012), of which trolling is a form. In a face-valuing culture (Leung & Cohen, 2011), where rejection avoidance is a major part of social interactions, extant literature would suggest that victims of trolling are likely to remain silent and evasive to preserve everyone's face (e.g., Hashimoto & Yamagishi, 2013). In an honour-valuing culture, however, it is typically believed that any insult requires retaliation to regain lost honour (Leung & Cohen, 2011), and so if a troll is perceived as offensive, we could expect victims reciprocate the aggression (e.g., Cohen, 1998).

Although culture's exact impact is seldom examined in detail when it comes to trolling (de Seta, 2013), aggression literature would suggest that one variable is particularly key: reputation (see Harinck, Shafa, Ellemers, & Beersma, 2013; Howell, Buckner, & Weeks, 2015; Uksul & Over, 2014). Depending on how a person's culture conceives of reputation (Leung & Cohen, 2011) and relationships with others (Markus & Kitayama, 1991), trolling victims may reciprocate aggression when offended (e.g., Cohen, Nisbett, Bowdle, &

Schwarz, 1996), or shy away from it, choosing instead to withdraw (e.g., Ma & Bellmore, 2016). Depending on the specific culture's beliefs about appropriate behaviours and responses, their exact strategy within the bounds of approach and avoidance should theoretically also change (Hashimoto & Yamagishi, 2013; Howell et al., 2015), being more aggressive or more reconciliatory, more direct or more subtle. Chapter 4 explores how countries with different norms and values react to various forms of trolling by directly exposing participants of differing backgrounds to overt and covert trolling, making them take center stage as victims in trolling interactions. In this way, we aim to experimentally determine how culture and trolling type impact victim reactions within trolling interactions.

However, we cannot neglect the context that the troll, bystanders, and victims all share during the trolling interaction: the anonymous online context. According to the SIDE model, when a group begins a conflict in an anonymous or pseudonymous context like an online game, the group will naturally split into two parties – for or against whatever the heart of the conflict is – and their actions and emotions will become increasingly polarized to match that of their side of the conflict (Postmes, Spears, & Lea, 1998). This is because people can act with impunity when anonymous, and are therefore free to act more outrageously or intensely online than they could offline. Translated to an in-game trolling situation, SIDE theory would posit that we can expect bystanders to side with either the troll or the victim and begin to attack the other camp of bystanders. We will look more at this in Chapter 5 when I put participants into a trolling situation in the role of a bystander, exploring how bystanders contribute to trolling interactions and how they experience being a witness to trolling behaviour. Combined with the victim's perspective presented in Chapter 4, this study gives trolling researchers a deeper understanding of trolling interactions beyond the influence of the troll alone.

Dissertation Outline

The global aim of the present dissertation is to explore the fundamental questions of trolling from an interdisciplinary perspective, determining a) what trolling is, b) why people do it, and c) who helps and who hinders trolling in online games. Each chapter also has a specific focus: the perspective of the troll (Chapter 2), a bird's eye view of the full interaction (Chapter 3), the victim's perspective (Chapter 4), and the bystander's perspective (Chapter 5). Chapter 2 aims examines troll's own perspective on trolling behaviour, something that was lacking in the majority of existing studies published at the time of its inception. More specifically, it aims to determine a) what trolls consider trolling, b) what motivates trolls to troll, and c) how the online community either encourages or discourages trolling practices. To uncover the answers to these questions, we conduct semi-structured interviews via Skype with over twenty self-confessed trolls. Participants were asked questions across a few broad themes: a) their experiences as a bystander of trolling, b) their experiences as a victim of trolling, c) their experiences as a perpetrator of trolling, and d) their opinion of the wider gaming community's role. These stories allow us to see trolls from the moment they learned what trolling was, to when they became perpetrators.

Chapter 3 takes a broader perspective and examines complete, authentic trolling interactions. For this, I analyzed a publicly available dataset comprising over 10,000 reported incidents of trolling in the popular online game League of Legends. These reports include game and player statistics for the game in which the troll was most recently reported, as well as the chat log from said game in which the troll was reported by other players in-game for trolling. After an extensive data cleaning procedure, the chat log data undergoes two separate analyses: structural topic modelling (STM), and a traditional dictionary-based content analysis. By conducting these two analyses across datasets split by actors, we are able to see which features from literature overlap between actors or can be used to differentiate actors

from one another. This allows us to simultaneously evaluate the descriptive contributions of different fields within trolling literature and apply them to different actors in actual trolling interactions to see what about what they say makes a troll a troll.

In Chapter 4, we shift our emphasis back to the individual perspective, focusing on the victims of trolling in particular to examine how their reactions shape trolling interactions. More specifically, this study aims to determine how offline culture can affect victims' intentions toward trolls, as well as their emotional and behavioural responses to two distinct types of trolling: flaming and ostracism. I contrast these two types of trolling for two primary reasons: a) both are prevalent online (see McCosker, 2014 and Williams, Cheung, & Choi, 2000 for examples), and b) they are distinct in their overtness, with flaming being highly abrasive and confrontational, while ostracism is much more subtle. We also contrasted three participants from three countries: the Netherlands, Taiwan, and Pakistan. These countries were selected because they are likely to differ in terms of how they conceive of reputation: as a matter of face, honour, or dignity (see Leung & Cohen, 2011). We expected that this may change how they react to more subtle and more overt form of aggression. Participants in this study were thrust into the victim role of a trolling interaction as either a local (ingroup) or a minority group member (outgroup) flames or ostracizes them while a bystander watches.

Chapter 5 brings the perspective of bystanders to the forefront using a similar paradigm to that used in Chapter 4, only instead of placing participants in the role of victim, we turn them into bystanders. The goal of this final study is to determine what makes gamers decide to intervene when witnessing a trolling interaction, as extant literature suggests that this decision can have a major impact on trolling interactions as a whole (e.g., Herring et al., 2002). The particular intervention that interests me in this study is reporting the troll to an authority figure, as this is considered one of the most effective strategies for deterring future trolling, and also one of the least-used (see Chapter 2). To this end, another experiment was

conducted reminiscent of that covered in Chapter 4, but placing participants in the role of the bystander instead of that of the victim. The study provides a glimpse into the experience of being a bystander and witnessing trolling, which is typically subtler an offense than most bystander studies employ (see Darley & Latane, 1978). The study as a whole also opens the door to new opportunities for non-automated intervention research in trolling as a field.

References

Barlett, C.P., Gentile, D.A., Anderson, C.A., Suzuki, K., Sakamoto, A., Yamaoka, A., & Katsura, R. (2014). Cross-cultural differences in cyberbullying behavior: A short-term longitudinal study. *Journal of Cross-Cultural Psychology, 45*, 300-313. DOI: 10.1177/0022022113504622

Bergtau. (2014, March). Re: Pranking vs. trolling: What is acceptable? *Day9tv*. Retrieved from https://day9.tv/d/SteppeLively/pranks-vs-trolling-what-is-acceptable/

Blackburn, J., & Kwak, H. (2014). *STFU NOOB! Predicting crowdsourced decisions on toxic behavior in online games*. Paper presented at the 23rd International World Wide Web Conference (WWW), Seoul, South Korea.

Blazyte, A. (2019, February). Gaming in China – Statistics and facts. *Statista*. https://www.statista.com/topics/4642/gaming-in-china/

Bresnahan, M.J., Shearman, S.M., Lee, S.Y., Ohashi, R., & Mosher, D. (2002). Personal and cultural differences in responding to criticism in three countries. *Asian Journal of Social Psychology, 5*, 93-105. DOI: 10.1111/1467-839X.00097

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

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017, February). *Anyone can become a troll: Causes of trolling behavior in online discussions*. Paper presented at the 20th conference on Computer-supported cooperative work and social computing, Portland, Oregon, United States.

Chesney, T., Coyne, I., Logan, B., & Madden, N. (2009). Griefing in virtual worlds: Causes, casualties and coping strategies. *Information Systems Journal*, *19*, 525-548. DOI: 10.1111/j.1365-2575.2009.00330.x

Cocomello, M. (2016, October). How to be a video game troll. *MyGaming*. Retrieved from https://mygaming.co.za/news/pc/106765-how-to-be-a-video-game-troll.html

Cohen, D. (1998). Culture, social organization, and patterns of violence. *Journal of Personality and Social Psychology*, *75*, 408-419. DOI: 10.1037/0022-3514.75.2.408

Cohen, D., Nisbett, R.E., Bowdle, B.F., & Schwarz, N. (1996). Insult, aggression, and the southern culture of honor: An "experimental ethnography". *Journal of Personality and Social Psychology*, *76*, 945-960. DOI: 10.1037//0022-3514.70.5.945

Cook, C., Conijn, R., Antheunis, M., & Schaafsma, J. (2019). For whom the gamer trolls: A study of trolling interactions in the online gaming context. *Journal of Computer-Mediated Communication, 24*, 293-318. DOI: 10.1093/jcmc/zmz014

Cook, C., Schaafsma, J., & Antheunis, M. (2018). Under the bridge: An in-depth examination of online trolling in the gaming context. *New Media & Society*, *20*, 3323-3340. DOI: 10.1177/1461444817748578

Donath, J. (1999). Identity and deception in the virtual community. In M.A. Smith and P. Kollock (Eds.), *Communities in Cyberspace* (pp. 29-59). London, Routledge.

Dutch Games Association. (2019). A closer look at the Dutch games industry. *Dutch Games Association*. Retrieved from https://dutchgamesassociation.nl/applied/closer-look-dutch-games-industry/

Entertainment Software Association. (2019). 2019 essential facts about the computer and video game industry. Retrieved from https://www.theesa.com/wp-content/uploads/2019/05/ESA_Essential_facts_2019_final.pdf

Fichman, P, & Sanfilippo, M.R. (2016). *Online trolling and its perpetrators: Under the cyberbridge*. Lanham, MD: Rowman & Littlefield.

Fox, F., Gilbert, M., & Tang, W.Y. (2018). Player experiences in a massively multiplayer online game: A diary study of performance, motivation, and social interaction. *New Media & Society*, *20*, 4056-4073. DOI: 10.1177/1461444818767102

Graham, E. (2019). Boundary maintenance and the origins of trolling. *New Media & Society*, *21*, 2029-2047. DOI: 10.1177/1461444819837561

Hardaker, C. (2010). Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research*, *6*, 215-242. DOI: 10.1515/JPLR.2010.011

Harinck, F., Shafa, S., Ellemers, N., & Beersma, B. (2013). The good news about honor culture: The preference for cooperative conflict management in the absence of insults. *Negotiation and Conflict Management Research*, *6*, 67-78. DOI: 10.1111/ncmr.12007

Hashimoto, H., & Yamagishi, T. (2013). Two faces of interdependence: Harmony seeking and rejection avoidance. *Asian Journal of Social Psychology*, *16*, 142-151. DOI: 10.1111/ajsp.12022

Henik, E. (2015). Understanding whistle-blowing: A set-theoretic approach. *Journal of Business Research*, *68*, 442-450. DOI: 10.1016/j.jbusres.2014.06.004

Herring, S., Job-Sluder, K., Scheckler, R., & Barab, S. (2002). Searching for safety online: Managing "trolling" in a feminist forum. *The Information Society, 18*, 371-384. DOI: 10.1080/01972240290108186

Howell, A.N., Buckner, J.D., & Weeks, J.W. (2015). Culture of honour theory and social anxiety: Cross-regional and sex differences in relationships among honour-concerns, social anxiety and reactive aggression. *Cognition and Emotion*, *29*, 568-577. DOI: 10.1080/02699931.2014.922055

Johnson, N.A., Cooper, R.B., & Chin, W.W. (2009). Anger and flaming in computer-mediated negotiation among strangers. *Decision Support Systems, 46*, 660-672. DOI: 10.1016/j.dss.2008.10.008

Kerr, E., & Lee, C.A.L. (2019). Trolls maintained: Baiting technological infrastructures of informational justice. *Information, Communication & Society*, 1-18. DOI: 10.17645/mac.v7i4.2347

Kwak, H., Blackburn, J., & Han, S. (2015). *Exploring cyberbullying and other toxic behavior in team competition games*. Paper presented at the CHI: Crossings conference, Seoul, South Korea.

Leung, A.K.-Y., & Cohen, D. (2011). Within- and between-culture variation: Individual differences and the cultural logics of honor, face, and dignity cultures. *Journal of Personality and Social Psychology, 100*, 507-526. DOI: 10.1037/a0022151

Li, Q. (2008). A cross-cultural comparison of adolescents' experience related to cyberbullying. *Educational Research, 50*, 223-234. DOI: 10.1080/00131880802309333

Ma, T.-L., & Bellmore, A. (2016). Early adolescents' responses upon witnessing peer victimization: A cross-culture comparison between students in Taiwan and the United States. *International Journal of Developmental Science, 10,* 33-42. DOI: 10.3233/DEV-150176

Maltby, J., Day, L., Hatcher, R.M., Tazzyman, S., Flowe, H.D., Palmer, E.J., … Cutts, K. (2015). Implicit theories of online trolling: Evidence that attention-seeking conceptions are associated with increased psychological resilience. *British Journal of Psychology, 107,* 1-19. DOI: 10.1111/bjop.12154

Markus, H.R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98,* 224-253. DOI: 10.1037/0033-295X.98.2.224

Massanari, A. (2017). #Gamergate and the Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society, 19*, 329-346. DOI: 10.1177/1461444815608807

McCosker, A. (2014). Trolling as provocation: YouTube's agonistic publics. *Convergence: The International Journal of Research into New Media Technologies, 20,* 201-217. doi: 10.1177/1354856513501413

Newzoo. (2018, August). Japan games market 2018. *Newzoo*. Retrieved from https://newzoo.com/insights/infographics/japan-games-market-2018/

van Osch, Y., Breugelmans, S.M., Zeelenberg, M., & Bölük, P. (2013). A different kind of honor culture: Family honor and aggression in Turks. *Group Processes and Intergroup Relations, 16,* 334-344. DOI: 10.1177/1368430212467475

Park, H.S., Levine, T.R., Weber, R., Lee, H.E., Terra, L.I., Botero, I.C., … Wilson, M.S. (2012). Individual and cultural variations in direct communication style. *International Journal of Intercultural Relations, 36,* 179-187. DOI: 10.1016/j.ijintrel.2011.12.010

Phillips, W. (2013, March). Ethnography of trolling: Workarounds, discipline jumping & ethical pitfalls (3 of 3). *Ethnography Matters*. Retrieved from https://ethnographymatters.net/blog/2013/03/05/ethnography-of-trolling-workarounds-discipline-jumping-ethical-pitfalls-3-of-3/

Phillips, W. (2011). Meet the trolls. *Index on Censorship*, *40*, 68-76. DOI: 10.1177/0306422011409641

Phillips, W. (2016). *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. Boston, MA: MIT Press.

Postmes, T., Spears, R., & Lea, M. (1998). Breaching or building social boundaries? SIDE-effects of computer-mediated communication. *Communication Research, 25,* 689-715. DOI: 10.1177/009365098025006006

Pozzoli, T., Ang, R.P., & Gini, G. (2012). Bystanders' reactions to bullying: A cross-cultural analysis of personal correlates among Italian and Singaporean students. *Social Development, 21*, 686-703. DOI: 10.1111/j.1467-9507.2011.00651.x

Rodriguez Mosquera, P.M., Fischer, A.H., Manstead, A.S.R., & Zaalberg, R. (2008). Attack, disapproval, or withdrawal? The role of honour in anger and shame responses to being insulted. *Cognition and Emotion, 22*, 1471-1498. DOI: 10.1080/02699930701822272

de Seta, G. (2013). Spraying, fishing, looking for trouble: The Chinese internet and a critical perspective on the concept of trolling. *The Fibreculture Journal, 1*, 301-318.

Retrieved from http://fibreculturejournal.org/wp-content/pdfs/FCJ-167Gabriele%20de%20Seta.pdf

Shachaf, P., & Hara, N. (2010). Beyond vandalism: Wikipedia trolls. *Journal of Information Science, 36*, 357-370. DOI: 10.1177/0165551510365390

Stein, J. (2016, August). Why we're losing the internet to the culture of hate. *Time*. Retrieved from https://time.com/4457110/internet-trolls/

Sue, J., & Mintegral. (2019, April). South Korea is a mobile gaming haven – even for Western studios. *Venture Beat*. Retrieved from https://venturebeat.com/2019/04/28/south-korea-is-a-mobile-gaming-haven-even-for-western-studios/

Suler, J.R., & Phillips, W.L. (1998). The bad boys of cyberspace: Deviant behavior in multimedia chat community. *CyberPsychology & Behavior*, *1*, 275-294. DOI: 10.1089/cpb.1998.1.275

Thacker, S., & Griffiths, M.D. (2012). An exploratory study of trolling in online video gaming. *International Journal of Cyber Behavior, Psychology and Learning*, *2*, 17-33. DOI: 10.4018/ijcbpl.2012100102

Uskul, A.K., & Over, H. (2014). Responses to social exclusion in cultural context: Evidence from farming and herding communities. *Journal of Personality and Social Psychology*, *106*, 752-771. DOI: 10.1037/a0035810

Williams, K.D., Cheung, C.K.T., & Choi, W. (2000). Cyberostracism: Effects of being ignored over the internet. *Journal of Personality and Social Psychology*, *79*, 748-762. doi: I0.1037//O022-3514.79.5.748

Chapter 2: Under the Bridge

An in-depth examination of online trolling in the gaming context

Trolling is a subject of apparent academic confusion; the few studies conducted thus far yielded a variety of conflicting definitions regarding what constitutes trolling behaviour and little information regarding trolling motivations. In order to shed further light on this phenomenon, the present study aimed to (1) determine which behaviours actual trolls consider as trolling, (2) explore the motivations behind trolling, and (3) examine the online community's response to trolling as perceived by the troll. After performing semi-structured interviews with 22 self-confessed trolls, we found that there is a variety of behaviours trolls consider trolling which can now be put in clear categories based on target and method. Three key motivations to troll emerged: personal enjoyment, revenge, and thrill-seeking. Trolling also appears to be a cyclical, self-perpetuating phenomenon enabled by the online community at large. Theoretical implications for future trolling research are also discussed.

This chapter is based on the following article:

Cook, C., Schaafsma, J., & Antheunis, M. (2018). Under the bridge: An in-depth examination of online trolling in the gaming context. *New Media & Society*, *20*, 3323-3340. doi: 10.1177/1461444817748578

Introduction

In the world of online gaming, undesirable behaviour is commonplace. Players will kill teammates, verbally abuse their peers, and misdirect new community members, spreading chaos and disorder (see Riot Games, 2015). These people are called 'trolls' and their behaviour 'trolling'. However, despite its prevalence in cyberspace, trolling as a subject of academic study is a confusing space, with different researchers using different criteria to describe the same phenomenon. This is likely due to the fact that it is such a new field of study: existing studies are few and far between, and nearly all of them have been atheoretical due to a lack of empirical basis upon which to build any theories (Herring et al., 2002; Shachaf and Hara, 2010; Thacker and Griffiths, 2012). Some researchers treat any deceptive action online as trolling (Buckels et al., 2014), while deception is not always required by other researchers (Fichman and Sanfilippo, 2014). Other negative behaviours with a perceived hostile intent are also sometimes grouped into trolling, while other researchers treat them as separate phenomena, such as griefing and flaming (Coyne et al., 2009; O'Sullivan and Flanagin, 2003; Thacker and Griffiths, 2012). With the concept of trolling in this fractured state, it is difficult at best to determine what to analyse when examining online behaviour in or out of game.

In addition to what trolls do, limited research has been conducted into the motivations and goals behind their actions – why and who they choose to troll and what gratifies them. Those that do address the community itself only indirectly (Herring et al., 2002; Luzón, 2011), examine only the victims' or bystanders' perspective (Maltby and et al, 2015; Shachaf and Hara, 2010), or lack an in-depth interview method (Buckels et al., 2014; Thacker and Griffiths, 2012). Finally, there is almost no available research pertaining to the influence the online community might have on trolling, despite multiple studies suggesting that the actions and attitudes of other netizens can have an important impact on both on- and offline

behaviour (Ridout and Campbell, 2014; Whitty and Carr, 2006; Young and Jordan, 2013). The present study aims to rectify these research gaps by moving past the survey method and instead performing in-depth interviews with actual trolls to (1) determine which behaviours actual trolls consider as trolling, (2) explore the motivations behind trolling, and (3) examine the online community's response to trolling as perceived by the troll.

Trolling Behaviour

As previously mentioned, there is a lack of academic consensus on the subject of trolling behaviour. Most researchers agree that trolling can fall into two large categories of verbal and behavioural trolling, with behavioural trolling being largely relegated to the gaming sphere. Beyond this basic agreement, discrepancies abound. A variety of negative statements, such as personal insults and exclusion tactics have been considered 'trolling' by some researchers (Herring et al., 2002; Luzón, 2011; Shachaf and Hara, 2010; Thacker and Griffiths, 2012), but other researchers will treat some of these options as separate from trolling (e.g. Alonzo and Aiken, 2004; Douglas and McGarty, 2001; Hardaker, 2010). Hostile intent is meant to be one of the few underlying threads that connect these verbal behaviours (Buckels et al., 2014; Fichman and Sanfilippo, 2014), but in the case of false ignorance (pretending to be ignorant of game mechanics or lore), the target of the trolling is the troll – they do not attack anyone else (Thacker and Griffiths, 2012). In terms of behavioural trolling, the examples most commonly given in the few studies that address this topic are team-killing and team-blocking, in which the troll either kills off other members of their team in-game or hinders their progress in completing objectives some other way. There are, however, other kinds of behavioural trolling that are not covered in the literature: feeding (allowing oneself to be killed by the opposing team to disadvantage one's own team and advantaging the opponents), spamming, and going 'AFK', or 'away from keyboard' (Riot Games, 2015).

Each of these behaviours consists of the abuse of a game mechanic that results in a disadvantage for opponents or teammates in-game.

In the past, researchers have often classified behaviours as trolling either based on reported incidents from community members (see Shachaf and Hara, 2010) or via observations of interactions in natural online settings and labelling parts of these interactions as trolling (see Suler and Phillips, 1998). This more bottom-up approach is both valid and useful in directing future trolling research. It remains, however, incomplete without its top-down counterpart. Although studies exist in which trolls are consulted, these studies use a survey methodology (see Buckels et al., 2014). This is a reasonable approach, but it is less exploratory, thus potentially missing phenomenological insights. The present study will fill this gap by talking directly to trolls, thus giving a community insider's perspective on which behaviours should be classified as trolling, in which environments, and under which circumstances. In this way, we can solidify our understanding of trolling on a variety of levels and complete our picture of trolling behaviour.

Trolling Motivations

In addition to determining which kinds of behaviours are considered trolling, we also aim to uncover the intent behind these questionable behaviours, as this element of intent appears key in the categorization of trolling as successful or failed. Typically, the success of a troll is determined by (1) examining second-hand accounts for emotive bystander or victim responses (see Herring et al., 2002) or (2) talking to victims or bystanders directly to see if a given behaviour had negative effects (see Fichman and Sanfilippo, 2014). However, this limits the researcher to looking at the interaction from a maximum of two perspectives: victim or bystander. Trolls as a population of interest allow for the unique research opportunity of speaking to one person from all perspectives in a trolling situation – the victim, the bystander, and the perpetrator. This allows us to ask questions from several angles

that are simply not possible with a different population, thus adding a new level of profundity to our understanding of trolling intention.

Integral to the question of intent, however, is that of motivation – what precedes the hostile intent and brings forth action? Although some studies have begun to explore various possible motives, these studies are few in number (Buckels et al., 2014; Shachaf and Hara, 2010; Thacker and Griffiths, 2012). More importantly, however, is their apparent homogeneity; in nearly all of these studies, trolls are presented as uniformly hostile and antisocial (Fichman and Sanfilippo, 2014; Herring et al., 2002; Shachaf and Hara, 2010). For example, Buckels and colleagues (2014) took a personality psychology approach to examining trolling antecedents. They found that trolling behaviour correlated positively with three out of the four components of the Dark Tetrad: sadism, psychopathy, and Machiavellianism. Because sadism correlated the strongest with trolling execution, Buckels et al. (2014) concluded that online trolling seems to be 'an Internet manifestation of everyday sadism' (p. 1). This conclusion effectively reduces trolling to a single, personality-driven cause. If trolls are more than a personality typology, however, it seems unlikely that they are all consistently antisocial. In addition to this, previous research has shown that not all trolling is antisocial; false ignorance, for example, does not necessitate harm done (Thacker and Griffiths, 2012). Given the variety of documented trolling behaviours (see Suler and Phillips, 1998), it seems much more likely that different motivations guide different behaviours in different situations. Making this distinction is critical for the field's advancement, as the fields of personality and community dynamics differ considerably. The present study will address this question of motivational homogeneity or heterogeneity by asking trolls directly what they intend by what they do and what triggers the problematic behaviours.

Role of the Community

Thus far, we have addressed two of our primary aims: determining which behaviours constitute trolling, and exploring the various motivations behind said behaviours. Although academic forays into trolling behaviour and motivation have been limited in number, even fewer studies address the role of context and community in the phenomenon. Is trolling in fact normative in the online community? Traditional psychology and sociology suggest that norms are formed and solidified over time (MacNeil and Sherif, 1976) and are largely based on three key factors: (1) inclinations (how a person thinks they should act), (2) regulatory interests (how that person thinks others should act), and (3) enforcement resources (the capacity of the group to enforce the rules; Heckathorn, 1988). These three factors interact over time to create norms in various settings, from experimental studies (Martin et al., 1974) to large-scale cultural conceptions of justice (Stolte, 1987). Thus, the question of trolling's deviance or normativity should be determined by examining these three factors. If trolling is deviant, as postulated by Fichman and Sanfilippo (2014), then trolls should be statistically abnormal in their inclinations, and the community should be essentially uniform in their regulatory interests, with medium to high levels of enforcement resources.

However, norm formation online is a little-studied topic, and to our knowledge, it has not been examined in the online gaming context prior to the present study. Thus, to begin to test the foundations of Fichman and Sanfilippo (2014) assertion, we are required to examine the gaming community's online presence directly. For example, by looking at the gaming community on YouTube, we can see that regulatory interests of the group appear divided. Trolling channels seem to be a popular form of entertainment on YouTube, with one such channel alone, videogamedunkey, having over 3,000,000 subscribers at the time of writing. By searching 'trolling' in YouTube's video search function, over 20,000,000 hits are produced in seconds (YouTube, 2017). Yet, on the fora of popular online game League of

Legends, there is an entire section dedicated to player behaviour. Its opening page is filled with questions such as ‘Riot why can people that do this get away with it’ and ‘Why is a player with a history of trolling not banned?’ (Riot Games, 2017). This latter example also suggests a lack of power on the part of the gaming community to regulate in-game behaviour, as the enforcement resources lie with the game administrators. Players are free to report, but the final say in terms of punishment belongs to the company. All of this gives reason to doubt the deviance of trolling within the community, but it remains anecdotal evidence. Here again the multiple perspectives trolls afford as online gamers will enable us to better answer this question on the community’s understanding and perception of trolling behaviour, thus allowing for the first scientific examination of these norms.

Method

Data Collection

For this study, in-depth interviews were conducted with self-confessed trolls, meaning they are aware of their trolling history and their continuance in the behaviour. Participants were allowed to have a history of trolling in any type of game, be it PC, console, or otherwise, in any genre. This was decided in order to reach the widest audience and have the most variety possible in our sample due to the exploratory nature of the study, and previous reports that trolls are an extremely difficult population to reach (see Shachaf and Hara, 2010). This self-identification as a troll was our primary concern. The only other participation criterion was age: participants must be 19 years of age or older due to ethical constraints. All participants were recruited using a combination of web-based advertisements, paper flyers, network sampling, and snowball techniques.

A total of 22 semi-structured interviews were conducted via Skype with a native-English speaking researcher (the first author) between February 2016 and May 2016. Three continents were represented in this sample, with 68% of participants living in North America

(10), 36% living in Europe (8), and 18% of participants living in South America (4). Of these participants, three identified as having a differing country of origin: one Indian participant and one Colombian participant were living in the Netherlands at the time of the study, and one South Korean participant was living in Canada. All other participants were residing in their home country. Participants were on average 23.6 years old (standard deviation [*SD*] = 2.4) and all had at least a high school diploma or its equivalent, while 32% had also completed some level of post-secondary education (one college certificate, four bachelor's degrees, two master's degrees). Participants had been gaming from 3.5 to 23 years, with an average of 14 years overall (*SD* = 5.78). Only 9% of the sample was female (2), which is a common finding in the extant literature (Buckels et al., 2014; Thacker and Griffiths, 2012), and prevented us from making gender comparisons. By the 22nd interview, saturation had been reached; irrespective of cultural background, stories were remarkably similar. Interviews lasted between 24 and 90 minutes (*M* = 51.06, *SD* = 15.42).

Data Analysis

Interviews were transcribed and it was these transcriptions that were analysed. Throughout this process, common themes or keywords were taken in note. Keywords were chosen to reflect themes present in the literature, such as trolling's apparent inherent negative nature (see Buckels et al., 2014), or for their frequency of use in question responses. At this point, questions were arranged by the three aforementioned themes (trolling behaviour, trolling motivation, role of the community) and a codebook was created. The primary researcher then coded the interview transcripts using the codebook. After this, a second coder experienced in gaming and game terminology was given the codebook and the interview transcripts and asked to code them a second time. At the initial coding, inter-rater agreement was at 63%. Recognizing that this was low for exploratory research (see Lombard et al., 2002), both coders met to discuss differences in their respective codes. After this discussion,

coding was adjusted to reflect what they had agreed upon (e.g. if a game glitch was exploited, it was considered contrary play). It was the final coding consensus that was used to rank the categories in frequency of use.

Results

Trolling Behaviour

In the questions regarding trolling behaviours, our sample described their own personal definition of trolling, as well as the various kinds of trolling they see exhibited in-game. They also expounded upon their own trolling behaviours.

Trolling Definitions

Consistent with the extant literature, trolls themselves also give a variety of different trolling definitions. However, these definitions can be generally split into three categories based on the elements of trolling they stressed: (1) attack, (2) sensation-seeking, and (3) interaction-seeking. These elements are not mutually exclusive. There can be elements of sensation-seeking within the interaction-seeking group or vice-versa, for example. However, each definition did have a primary stress, and it was by this emphasis that they were categorized.

According to the participants who stressed the element of attack in their definitions, trolling is a direct attack on the other players' enjoyment of the game or gameplay. This type of definition was the most common seen in the sample, and tends to be the view of trolling commonly presented in the extant literature, gaming-focused or otherwise (Buckels et al., 2014; Herring et al., 2002; Shachaf and Hara, 2010). Participants holding this viewpoint called trolling 'ruining gameplay for other people' (P15, 24, female) and described it as 'intentional loss, or people playing with the intent to piss other players off' (P12, 24, male). In this definition category, trolling is presented as purely anti-social and antagonistic.

Participants emphasizing sensation-seeking in their definitions painted a more asocial picture of trolling: it is neither inherently good nor bad, but simply a behaviour which leads to enjoyable consequences for the troll. Participants described it as 'the creation of drama' (P22, 20, male) and as a way to get attention. The victim's reaction figured heavily as a source of thrill or enjoyment in these definitions, as it is this drama that satisfies the troll. Typically, the more outrageous the reaction, the better (P11, 23, male; P20, 23, male). This definition category is typified by its hedonistic flair and its emphasis on the other players as a source of pleasure, be it sadistic or otherwise.

Finally, participants who emphasized interaction-seeking defined trolling as an unorthodox method of communication designed to make players get involved in both the conversation and the game. These definitions can present trolling as either prosocial or asocial, but never antisocial. The following passage is typical of a definition in this category:

> For me, [trolling is] mostly when somebody is doing well but then doesn't … hmm … want to play the game just for the game but more for like … hm … interaction purposes. Trolls generally interact a lot. So we like participation from the other side too. It gets really boring when you are the only one trolling. It has to be a few trolls to be fun. (P4, 27, male)

Like the sensation-seeking category, this definition type also emphasizes other players and their reactions to trolling. However, the desired reaction is radically different between the two. Based on this definition, a positive response is much more highly valued than an outrageous one. The desire is not sensation, but friendship. In fact, some participants even argue that trolling between friends actually makes the game more enjoyable (P10, 23, male: 'Trolling with fun and friends, I think they laugh about it and it's ok'). Whether it occurs between friends or friends-to-be, this definition of trolling places particular importance on amusing, albeit unorthodox, interaction between players.

In-Game Behaviour

Based on the participants' reports, trolling behaviours can be largely divided into two groups: verbal trolling and behavioural trolling. Table 1 presents a summary of both types, their sub-types, and an explanation of these. Within the verbal category of trolling, trash-talking is the most commonly mentioned, and is typically comparable to what you might see in a traditional sports arena:

> Oh, so um, while the game is loading, at first you can write messages to each other. So they're like 'Yeah, we're gonna win! We're gonna defeat you!' Like sometimes just talk with their … like for example, 'I've won this many times!' For example with my cousin or my brother, we're like 'Yeah you're the worst! My army is so good. You'll always lose!' and that kind of stuff. (P16, 21, male)

However, more negative trash-talking is also present in the sample. This consists of direct and often unjustified criticism of another player on a personal or gameplay level, or in some more extreme cases, even insulting or degrading the player's family members (P3, 24, male 'Your mom is a B, like a B-word').

Table 1.

Trolling types and sub-divisions.

Trolling Type	**Explanation**
Verbal Trolling	Using a chat function in-game to troll another player.
Trash-talking	Putting down or making fun of other players
Flaming	Presenting emotionally-fueled or contrary statements with an instrumental purpose.
Misdirection	Spread false information among targeted or general players
Spamming	Repeating game-unrelated chat either textually or audibly in-game.
Inappropriate Roleplaying	Pretending you are a different person (non-game-related) to obtain some kind of specific reaction.
Behavioural Trolling	Using existing game mechanics to troll another player.
Inhibiting Team	Actively hampering your teammates' in their goals.
Contrary Play	Playing the game outside of what is intended by most players.
Aiding the Enemy	Disregarding strategic play to make it easier for the opposing team to win.

In terms of behavioural trolling, individual behaviours were grouped into the following three categories: (1) inhibiting your team, (2) contrary play, and (3) aiding the

enemy team (see Table 1 for descriptions). Although all three were reported regularly by participants, inhibiting your team was the most popular of the sample-described trolling methods. Essentially, team inhibition consists of playing the game with the goal of ruining the gameplay of your teammates. It can be as complicated as throwing your character in front of the attacks of friends in order to make it look like they are killing their own team members and incur a penalty, or as simple as blocking a path required for your teammates to reach their goals:

> Well um … for COD for instance, you could uh … corner-trap your teammates. They'll be just camping in the corner and then if you stand in front of them, they can't escape … and they get really mad. (P9, 22, male)

Examples of contrary play would be exploiting a game glitch to gain an advantage over other players, or pushing another player's avatar into a body of water, while aiding the enemy team could be broadcasting team positions or 'feeding'. Interestingly, all of these behaviours were also instrumental. If there was no goal or ulterior motive attached, for example, in a flaming or contrary play situation, then it was not classified as trolling. When referring to trolling, participants consistently included a goal, such as obtaining a reaction from others or distracting the enemy players. Instrumentality figured strongly as a key indicator of trolling classification in our sample.

Trolling Motivations

In the questions regarding trolling motivations, our sample described why they exhibit the aforementioned behaviours by clarifying their 'triggers' and associated goals – essentially, why they troll. Although goals and triggers may seem to overlap – both have clear connections to motivation – it is important to note that they were treated separately during the interviews. Triggers were referred to as a catalyst to begin trolling, while goals refer to the ultimate achievement desired by the troll. Despite some similarities in response, we consider these to be conceptually distinct.

Trolling Triggers

Despite some claims that it happened ‘randomly’, all members of the sample were able to identify one or more triggers that typically preceded their trolling. These events could be broadly categorized into the following three types: (1) social triggers, (2) internal triggers, and (3) circumstantial triggers.

Social. In our most popular and broad category, there was one clear forerunner for type of trigger: being trolled. Of all the social triggers listed, being trolled first was the single most popular reason to begin trolling:

> It sounds super silly, but more often than not, if I’m trolling it’s because I’ve been trolled, and it’s kinda in the hopes that they stop, and they realize that it’s super annoying, and that they wanna win the game, and that it’s way easier to win the game if there are 5 people participating instead of 3 people participating. So you know, I hope that it stops, or they at least leave me alone so that I can play my best game. (P15, 24, female)

Trolling appears to breed trolling, with the behaviour seemingly becoming a social contagion among gamers. Other social triggers typically involved noticing weakness in other players, either poor gameplay or general gullibility. However, these border on internal triggers, as the player has to be seeking signs of weakness or vulnerability in order to spot them quickly and act upon them.

Internal. Eleven members of the sample mentioned internal triggers as the catalyst to their trolling. In fact, with the exception of the two female trolls in the sample, all participants who mentioned more than one trigger mentioned at least one internal trigger. Of these internal triggers, two emerged as the most common overall: being ‘on tilt’, referring to a negative emotional state in which gameplay typically suffers, and boredom:

> R: Why did you start to troll yourself?
> P9: Personal enjoyment.
> R: Something to do?
> P9: Yeah, you get bored of the game sometimes, and sometimes it makes the game more exciting. (P22, male)

Those who troll due to boredom often seem to treat the trolling as a sort of meta-game. They see themselves as being 'beyond' the game, having seen all there is to see, and thus try to 'win' at trolling instead. Remarkably, the other primary internal trigger, being on tilt, appears to be primarily caused by consecutive loss:

> R: Yeah, so for you [trolling is], it's a reaction based on anger, it's lashing out in anger.
> P22: Yeah.
> R: And the two things that happen beforehand, it's losing, it's triggered by losing, or …
> P22: Yeah, but losing on a streak, like, like lots of times in a row and that just … pissed me off. (20, male)

Thus, both veteran players and newbies can fall victim to internal triggers: if you are experienced in the game, you risk boredom; and if you are brand new, you risk consecutive loss against more experienced players. It is worth noting, however, that trolling itself is also listed as a reason someone may be on tilt, or one of the factors playing into a loss-streak. This strongly suggests a vicious-circle-like element to trolling.

Circumstantial. Least-popular among the triggers are the circumstantial triggers – only five members of the sample even mentioned them. There are two different circumstances that the sample listed as potential trolling-triggers: the pre-game, and a winning start to a match. In most online games, there is a 'pre-game lobby' in which players select their characters or avatars. Players can also chat in this pre-game lobby. The pre-game seems to set the tone for the rest of the game, and it is where trolling initiates in games with circumstantially triggered trolls. Despite the few trolls who mentioned them, this seems to suggest that the earliest part of the game is a partial determinant of trolling. For at least some trolls, simply having the opportunity to communicate pre-game is enough to initiate trolling.

Trolling Goals

A variety of individual motivations emerged from the interviews, but these can be grouped into three broader categories in order of sample frequency: (1) personal enjoyment, (2) revenge, and (3) thrill-seeking.

Personal enjoyment. Based on the sample, trolls motivated by personal enjoyment can either derive their joy from the sheer pleasure of trolling itself, or can use trolling strategically to disable or weaken their opponents in-game and enjoy winning the game:

> It's kinda like a game, almost? It's the game within the actual game. Like, if you, if all five of you go in and you can manage to win with shields, it's like, even more fun than just normally winning the round by playing the game. (P14, 19, male)

Trolling thus becomes a meta-game, an added challenge that heightens the gaming experience. This goal is closely linked with the boredom trolling trigger, as most participants in this category mention being disenchanted with the actual game and wanting something more from the gameplay (P14, 19, male: 'Eventually you just get bored and you're like, let's like, let's make it fun, let's make it funny'.)

Revenge. Revenge-motivated trolls, who are uniformly people who have first been trolled themselves, seek either the misery and/or failure of the initial troll, or the reformation of said troll's behaviour by showing them how their behaviour affects others negatively:

> The only reason why I would troll is if someone does it to me. It's not that I would go do something like that because I really want a lane and I'm just going to insta-lock – that's not me. I troll someone if he trolls me, so it's just a response. (P8, 19, female)

The above participant falls into a previously unknown trolling archetype – the vigilante. These trolls prey exclusively on other trolls to 'give them a taste of their own medicine,' so to speak, in the hopes of either reforming them or scaring them away from the online community. Other revenge-motivated trolls are purely 'reactionary' (P10, 23, male: 'It's a response to something, and most the times, if it's something stupid, I react about it normal, but if they act stupid again, I might flame them or something'.), who troll instinctually when

trolled, as opposed to with a specific goal in mind. Both, however, are considered revenge-motivated due to the requirement of being trolled first before taking action themselves.

Thrill-seeking. Thrill-seeking trolls seem to most resemble the trolling depictions presented in the extant literature (Buckels et al., 2014; Fichman and Sanfilippo, 2014; Thacker and Griffiths, 2012). These trolls in our sample had a noted disregard for the potential impact of their behaviour, single-mindedly seeking the most outrageous reaction possible by any means necessary, verbal or behavioural:

> Uh, you annoy other people. The same thing – why would you play against other players? It's, well you can kill their avatar. If you do it in a way that works, that's kind of nice too, but if you can do it in a way that they get pissed off even more, that's even more fun. (P19, 27, male)

Like their enjoyment-seeking counterparts, these trolls also frequently cite fun as a key goal in their enterprises. However, there is an additional goal in this category that does not appear in the personal enjoyment sphere: satisfaction of curiosity. Several trolls in our sample treat trolling as a 'social experiment' (P3, 24, male) designed to 'gauge' (P1, 24, male) other players. They satisfy their curiosity by, in their mind, empirically testing their preconceptions of other players. In either case, thrill-seeking trolls appear to be the most aggressive of the troll types.

Role of the Community

In the final set of questions regarding the role of the community, our sample discussed the kinds of responses to trolling they see in the community, both when they are the perpetrator and when they are a simple bystander. They also examined their own thoughts regarding the online community's opinions of trolling.

Trolling Responses

When questioned regarding typical responses given by victims and bystanders to trolling, once more our sample gave a variety of answers. These were condensed into five

categories: rage – verbally expressing negative feelings towards the troll; ignore – taking no action related to the trolling; troll – either joining in the troll's victimizing or trolling the troll back; prevention – muting or reporting the troll; and participation – joining in the conversation surrounding the trolling without trolling or raging oneself (bystanders only). These categories were then ranked for victims and bystanders according to the frequency with which responses falling into these categories were mentioned by participants. For bystanders, participation was the most popular reaction, followed by ignoring the troll, trolling themselves, raging, and finally prevention. Victims show a similar but differing pattern, with rage as the most popular response, followed by ignoring the troll, trolling back, and finally prevention.

Thus, it seems as though victims most often become angry and respond in kind when trolled, further entrenching the cyclical nature of trolling. According to our sample, rage and trolling back are two of the top reactions to trolling among victims. As discussed previously, thrill-seeking trolls are motivated by strong reactions, rage (and flaming) included; thus, by raging, victims motivate thrill-seeking trolls to continue trolling, while trolling back further entrenches trolling behaviour in the community. Interestingly, prevention behaviours such as reporting the troll or muting them are the least popular response, despite the fact that four members of our sample list 'getting reported' as a negative consequence and deterrent of trolling. This seems to indicate that victims of trolling also tend to be enablers of trolling, reacting in such a way that trolling is encouraged, and not taking advantage of built-in systems like the mute button and reporting to prevent the behaviour from happening again in the future. That said, ignoring is still a relatively popular option, and being ignored is also listed by nine of our trolls as a negative response to trolling, meaning that it discourages the troll. Still, the vast majority of victims seem to engage in responses that encourage further trolling in the community.

Based on our sample, bystanders are just as likely to engage in trolling prevention as victims, and also have their share of enabling behaviours. Participation in the trolling conversation, which typically entails either defending the victim or trying to 'talk down' the troll, emerged as the most popular bystander response to trolling. This, although perhaps not as direct as with the victim, provides trolls with further reactions to their behaviour, motivating the thrill-seekers and potentially even the personal enjoyment-motivated trolls. Ignoring was also a popular option, suggesting that bystanders are not constantly engaging in enabling behaviour. Typical enablers, such as trolling back or raging, were less popular among bystanders than among victims. That said, preventative behaviours remained the least popular response to trolling, despite its aforementioned capacity to reduce trolling in online communities. Thus, it appears that although bystanders do not seem to react as strongly as victims to trolling situations, they are still prone to troll-enabling by neglecting preventative measures and providing additional reactions to the trolling behaviour.

Trolling: Normative?

Before addressing trolling's normativity, it is crucial to understand the degree of nuance presented in our trolls' responses. There were few if any cut and dry answers regarding the normativity of trolling. When asked what they think the gaming community feels about trolling, 19 members of the sample said that it was negatively perceived. Participants used terms like 'necessary evil' and 'guilty pleasure' when describing trolling from the community's perspective, indicating its negative nature. Other participants described it as 'toxic' and 'annoying', cementing it as a darker side of gaming. A few participants also mentioned that trolling had evolved into its current negative state, and that it was the new generation of trolls that was twisting its original purpose (P20, 23, male; P21, 24, male). According to one participant in particular, trolling started as a way to play mind games and trick other players in games, but now the term has come to include all negative behaviours

online (P20, 23, male). Trolling is thus exposed as a dynamic, evolving phenomenon. Participant 12 may have summed all of these thoughts up best when he described trolling as ‘a problem that will never go away’ (24, male).

This said, ten members of the sample also mentioned at some point in the interview that trolling was a part of gaming, an inextricable piece of the activity. In addition to this, some sample members cited YouTube channels as proof of trolling’s normalcy:

> Um … … …. well there’s a lot of Youtube channels dedicated toward trolling people, so I’d say a lot of people just think it’s funny. They enjoy it when it’s not them. But I would say, if they weren’t being trolled themselves, they could enjoy someone else being trolled. Then, everyone seems to like it. (P9, 22, male)

This suggests that trolling is not only normative in the community, but even celebrated by some of its members, trolls and everyday gamers alike. Other participants made the distinction that, among friends, trolling is completely acceptable, and that it only treads into negative territory when it takes place among strangers (P10, 23, male: ‘So … yeah, trolling for fun … it’s ok, but trolling to um, influence other people that you don’t know is not ok’.). As previously mentioned, there is a high degree of nuance in our trolls’ responses, and context seems to be an important factor.

All of these varying responses seem to suggest that trolling is neither normative, nor deviant, but rather somewhere in between. It is clear that trolling is considered a negative phenomenon, but it is also an expected phenomenon, a ‘rite of passage’ (P4, 27, male) within the online gaming community. Were trolling normative, there would be repercussions for not engaging in trolling behaviour. In other words, the regulatory interests of the group would be activated by non-trolls. However, were trolling deviant, it would not have such a high prevalence, nor would it be expected to the current degree it is in the community. Thus, we propose trolling as an a-normative phenomenon – neither deviant nor prescribed, but an

active part of the community and largely tolerated. Whether considered positive or negative, however, nearly the entire sample agreed that it was normal behaviour within the community.

Discussion

Conclusions and Implications

At this study's outset, we aimed to examine what constituted trolling behaviour, what kinds of motivations and goals were associated with trolling, and how the community impacts trolling. Prior to this study, there were multiple, occasionally contradictory, definitions of what constitutes trolling floating in academia. We were able to confirm that this trend extends into the trolls' world as well – definitions provided by our sample were varied, though they generally fell into a few key thematic categories: attack, sensation-seeking, and interaction-seeking. This suggests that trolling is not a uniform phenomenon, but rather an umbrella term for certain types of instrumental online interactions. Attack trolls want their victims' misery, interaction-seeking trolls want friendship or conversation, and thrill-seeking trolls want sensation for themselves. This finding of instrumentality in trolling contradicts what is suggested by Buckels et al. (2014) trolling definition: trolling is actually characterized by its instrumentality, and not a wanton nature. In fact, negative or controversial online behaviours that are not typically considered trolling are often categorized as trolling once they develop an instrumental purpose. Take the example of flaming, which O'Sullivan and Flanagin (2003) define as 'hostile and aggressive interactions via text-based computer mediated communication' (p. 69). We found that flaming only entered the realm of trolling if it was used specifically to obtain an outrageous reaction from victims in the pursuit of sensation or thrills. This is a crucial insight into trolls' perception of trolling, and this instrumentality should be taken into careful consideration in future studies.

There remained also the question of what constituted trolling. Once more, there were several examples that had been documented (see Herring et al., 2002 and Shachaf and Hara,

2010), but they had not been categorized in any systematic fashion and were typically deemed ‘trolling’ by either researchers or laymen. There was no previous research classifying trolling behaviour according to the actual perpetrators. Through our interviews, we were able to develop a clear classification system of trolling behaviour from the trolls’ perspective: verbal (trash-talking, flaming, misdirection, spamming, and inappropriate roleplaying) and behavioural (inhibiting your team, aiding the enemy team, and contrary play). We also found that these behaviours appear to be dispersed unevenly across generations of gamers, creating a generational gap between trolls. Veteran gamers take on a trickster archetype when they troll, and tend towards misdirection and subterfuge, while new and younger gamers go for a more abrasive approach, engaging in behaviours such as trash-talking and killing teammates. This has caused veteran gamers to renege on the term ‘troll’, as they perceive it to have a different meaning today. While it was once a badge of honour symbolizing their mastery of intellect and gameplay, it has turned into a sign of shame reviled by most gamers. This finding suggests a hierarchical aspect of the online gaming community, and particularly for trolls, only hinted at previously (see Thacker and Griffiths, 2012), and never before attributed to gaming experience and age.

In addition to the question of trolling behaviour, we also sought to uncover the motivations behind said behaviour. The present study allowed for an in-depth examination of trolling motivation, revealing a variety of possible motivations and goals, ranging from prosocial to antisocial, and including personal enjoyment, revenge, and thrill-seeking. In particular, the importance of personal enjoyment and boredom as presented by the extant literature (Buckels et al., 2014; Thacker and Griffiths, 2012) was confirmed, as these also emerged as some of our sample’s most popular motivations and catalysts to trolling. However, we also uncovered otherwise unknown motivations, such as interaction-seeking and looking for friendship via trolling. Another previously undocumented phenomenon also

emerged from the interviews: the ‘vigilante troll’ – trolls who target other trolls with the goal of reforming their behaviour or exacting revenge. Interestingly, the two women trolls that we interviewed fell into this trolling motivation type. Although not generalizable to all women trolls due to the small sample size, it is still an important phenomenon to note, specifically in the light of GamerGate, an online event in which gamers of all stripes banded together against what they perceived as invasive feminist rhetoric (Chess and Shaw, 2015; Massanari, 2015; Mortensen, 2016; Vermeulen et al., 2016). Much of the literature on this topic is in the feminist tradition (Chess and Shaw, 2015; Massanari, 2015; Parkin, 2014; Todd, 2015) and espouses that despite increased presence of women in the gaming community and gaming industry, what Chess and Shaw (2015) call a ‘hegemonic masculinity’ (p. 218) pervades online gaming. However, another line of research focuses on the emotions involved in GamerGate – specifically a sense of victimization and rage on the part of gamers (Mortensen, 2016). It is these feelings that Mortensen (2016) asserts led to the GamerGate scandal, with gamers organizing themselves as a community to stand up to what they saw as their persecutors. This idea of vigilante justice is remarkably similar to the motivations our female trolls displayed when trolling other, statistically male, trolls. Although this motivation is not relegated exclusively to female trolls, further research is required to determine gender’s true role in trolling.

These motivational findings also link in to the idea of trolling being fundamentally instrumental. Some have alleged (see Buckels et al., 2014) that trolling is an aimless pursuit. However, trolls consistently list multiple associated goals and motivations behind the behaviour when asked (i.e. personal enjoyment, revenge, reformation, fun, etc.). We found that friendships can be formed and cemented via trolling, and that vigilante trolls seek to reform or remove other trolls from their game-space. Trolls have goals, and these vary dramatically from troll to troll. We also identified several trolling catalyst-events, called

'triggers' here, for modelling purposes. Many of the trolling triggers identified in this study are social, and should be detectable via analysis of chat logs and the like. However, some of the most common triggers and motivations were internal, with boredom being of particular importance. Trolling researchers must be careful to include variables such as mood and state of mind in future studies to ensure that they are taken with proper consideration in any modelling or empirical testing.

In terms of community and its impact on trolling, a major implication from the present study is the fact that there is a trolling community at all. Were there no form of community, it would have been impossible to recruit using a network sampling technique. As it stands, trolls are aware of one another and are often connected to other trolls, forming at the very least a loose community. In fact, many members of our sample reported trolling more often in groups than alone. Historically, trolls have been treated largely as individuals (see Buckels et al., 2014). However, this finding opens the way for group-level analyses and the exploration of these trolling communities. In addition to uncovering a trolling community, however, we also explored the normativity of trolling in the overall gaming community. Yet again, our findings contradict those presented by the extant literature (Fichman and Sanfilippo, 2014); trolling is, in fact, a normal, expected event, sometimes even described as a rite of passage. No one escapes it, and it thus becomes a shared, common experience between gamers, cementing the community. It is also anormative – tolerated, but not encouraged. In order to be considered deviant, trolling must enter the realm of cyberbullying or cybercrime, typically by persistently targeting a single person or entity repeatedly or by breaking a written law. It is important to note, however, that we did not find that trolling is considered positive. It may be common and expected, but our sample was unanimous in saying that it is seldom an enjoyable experience to be the victim of a troll. Thus, it seems to trolls and gamers alike that trolling is normal, but negative.

However, what is perhaps the most important and novel finding from this study is that trolling appears to be a cyclical, self-perpetuating phenomenon – the community's role in trolling appears to be enabling and perpetuating behaviour. Every single member of our sample reported having at some point been a victim of trolling themselves. Based on their responses, the cycle is strongly reinforced by the community and its response to trolling. We found that victims will more often than not respond to trolls by trolling them back. Our own trolls frequently reported carrying negativity forward into future games, suggesting that these initial victim responses can easily translate into future trolling experiences in which the initial victim becomes the perpetrator. In addition to this, bystanders tend to fuel the flames by jumping into the conversation between troll and victim, giving the troll an even larger reaction. By contrast, both victims and bystanders are relatively unlikely to engage in preventative action, thus supporting trolling by both omission and commission, however indirectly. These findings together form the greater finding that bystanders and victims are, however unwittingly, complicit in the trolling process. This could have potential connections to cyberbullying, as this too is reportedly a cyclical phenomenon (Vandebosch and Van Cleemput, 2009). Other studies have suggested that internet identity and cyberbullying identity is fluid, with bystanders and victims and perpetrators all interchanging roles over time (Park et al., 2014). In either case, both phenomena appear to be self-perpetuating and negatively perceived.

Limitations and Future Directions

As with all studies, this one is not without its limitations. One such limitation is in terms of the sample. Due to ethical constraints, we were only able to interview participants aged 19 or over. However, many of our participants suggested that trolling is even more common in younger audiences, and that there is a generational gap between veteran and young trolls. Thus, it would have been ideal to have more trolls and a wider age variance in

our sample to better determine whether or not this allegation is accurate, or merely a perceptual bias on the part of veteran gamers. Future studies could contrast age groups, or perform a similar study targeting a different age range to explore this apparent trend in-depth; with a larger sample available, they could contrast cultures as well. We also interviewed trolls specifically. This in of itself is not problematic, and was in fact the goal of the study. However, we discussed not only their experiences trolling, but also those of the bystanders to and victims of their trolling. Once more, this provides a different perspective on trolling, but should be taken carefully, as perpetrators within community settings have been previously shown to misperceive said community's norms and values (Young and Weerman, 2013). Thus, although the trolls' perception is a novel finding, further studies are required to be sure that this evaluation of community norms is not a perceptual bias on the trolls' part.

In addition, there remain some outstanding questions regarding the homogeneity or heterogeneity of the trolling community. The present study examined gamers, but there are other places that trolls can practice their craft – social media websites, forums, comment sections and the like. The differences and similarities between trolls and trolling behaviours on these different platforms has yet to be compared and contrasted. Answering this question could even potentially explain the differing results found by trolling researchers thus far (Buckels et al., 2014; Fichman and Sanfilippo, 2014; Shachaf and Hara, 2010; Thacker and Griffiths, 2012) if it is due to platform differences. Culture may also come into play here. The present study examined many different cultural groups, touching on three continents. Due to the difficulty in finding trolls willing to be interviewed, this happened naturally in the recruitment process. However, these groups were too small to make a truly generalizable cultural comparison. By examining how different cultures troll and how people troll on different platforms, we can enrich our understanding of the phenomenon as a whole. The

present study is a foundation of things to come. There is still much to be done before trolling as a subject of academic study can reach its full potential.

References

Alonzo, M., & Aiken, M. (2004). Flaming in electronic communication. *Decision Support Systems*, *36*(3), 205–213. DOI: 10.1016/S0167-9236(02)00190-2

Barreto, M., & Ellemers, N. (2002). The impact of anonymity and group identification on progroup behavior in computer-mediated groups. *Small Group Research*, *33*(5), 590–610. DOI: 10.1177/104649602237680

Bohannon, L. S., Herbert, A. M., Pelz, J. B., & Rantanen, E. M. (2013). Eye contact and video-mediated communication: A review. *Displays*, *34*(2), 177–185. DOI: 10.1016/j.displa.2012.10.009

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

Chesney, T., Coyne, I., Logan, B., & Madden, N. (2009). Griefing in virtual worlds: Causes, casualties and coping strategies. *Information Systems Journal*, *19*(6), 525–548. DOI: 10.1111/j.1365-2575.2009.00330.x

Coyne, I., Chesney, T., Logan, B., & Madden, N. (2009). Griefing in a virtual community: An exploratory survey of Second Life residents. *Zeitschrift Für Psychologie/Journal of Psychology*, *217*(4), 214–221. DOI: 10.1027/0044-3409.217.4.214

Croes, E., Antheunis, M., Scouten, A., & Krahmer, E. (2016). Teasing apart the effect of visibility and physical co-presence to examine the effect of CMC on interpersonal attraction. *Computers in Human Behavior*, *55*, 468-467. DOI: 10.1016/j.chb.2015.09.037

Douglas, K. M., & McGarty, C. (2001). Identity ability and self-presentation: Computer-mediated communication and intergroup interaction. *British Journal of Psychology, 40*, 399–416. DOI: 10.1348/014466601164894

Fichman, P., & Sanfilippo, M. R. (2014). The bad boys and girls of cyberspace: How gender and context impact perception of and reaction to trolling. *Social Science Computer Review*, *33*(2), 163–180. DOI: 10.1177/0894439314533169

Hardaker, C. (2010). Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research, 6*(2), 215–242. DOI: 10.1515/JPLR.2010.011

Heckathorn, D. D. (1988). Collective sanctions and the creation of prisoner's dilemma. *The American Journal of Sociology*, *94*(3), 535–562. Retrieved from https://www.jstor.org/stable/pdf/2780253.pdf

Herring, S., Job-Sluder, K., Scheckler, R., & Barab, S. (2002). Searching for safety online: Managing "trolling" in a feminist forum. *The Information Society, 18*(5), 371–384. DOI: 10.1080/01972240290108186

Lea, M., & Spears, R. (1991). Computer-mediated communication, de-individuation and group decision-making. *International Journal of Man-Machine Studies*, *34*(2), 283–301. DOI: 10.1016/0020-7373(91)90045-9

Lea, M., Spears, R., & Groot, D. De. (2001). Knowing me, knowing you: Anonymity effects. *Personality and Social Psychology Bulletin*, *27*(5), 526–537. DOI: 10.1177/0146167201275002

Lombard, M., Snyder-Duch, J., & Campanella Bracken, C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*(4), 587-604. doi/10.1111/j.1468-2958.2002.tb00826.x/epdf

Luzón, M. J. (2011). "Interesting post, but I disagree": Social presence and antisocial behaviour in academic weblogs. *Applied Linguistics*, *32*(5), 517–540. DOI: 10.1093/applin/amr021

MacNeil, M. K., & Sherif, M. (1976). Norm change over subject generations as a function of arbitrariness of prescribed norms. *Journal of Personality and Social Psychology, 34*(5), 762–773. DOI: 10.1037//0022-3514.34.5.762

Maltby, J., Day, L., Hatcher, R. M., Tazzyman, S., Flowe, H. D., Palmer, E. J., … Cutts, K. (2015). Implicit theories of online trolling: Evidence that attention-seeking conceptions are associated with increased psychological resilience. *British Journal of Psychology*, *107*(3), 448–466. DOI: 10.1111/bjop.12154

Martin, J. D., Williams, J. S., & Gray, L. N. (1974). Norm formation and subsequent divergence: Replication and variation. *The Journal of Social Psychology*, *93*(2), 261–269. DOI: 10.1080/00224545.1974.9923160

O'Sullivan, P. B., & Flanagin, A. J. (2003). Reconceptualizing "flaming" and other problematic messages. *New Media & Society, 5*(1), 69–94. DOI: 10.1177/1461444803005001908

Park, S., Na, E. Y., & Kim, E. (2014). The relationship between online activities, netiquette and cyberbullying. *Children and Youth Services Review*, *42*, 74–81. DOI: 10.1016/j.childyouth.2014.04.002

Postmes, T., & Baym, N. (2005). Intergroup dimensions of internet. In J. Harwood & H. Giles (Eds.), *Intergroup communication: Multiple perspectives* (pp. 213-238). New York: Peter Lang Publishers.

Postmes, T., Spears, R., & Lea, M. (1998). Breaching or building social boundaries? SIDE-Effects of computer-mediated communication. *Communication Research, 25*(6), 689–715. DOI: 0803973233

Ridout, B., & Campbell, A. (2014). Using Facebook to deliver a social norm intervention to reduce problem drinking at university. *Drug and Alcohol Review, 33*(6), 667–673. DOI: 10.1111/dar.12141

Riot Games. (2015). Reporting a player. Retrieved January 21, 2016, from https://support.riotgames.com/hc/en-us/articles/201752884-Reporting-a-Player

Shachaf, P., & Hara, N. (2010). Beyond vandalism: Wikipedia trolls. *Journal of Information Science*, *36*(3), 357–370. DOI: 10.1177/0165551510365390

Spottswood, E.L., Walther, J.B., Holmstrom, A.J., & Ellison, N.B. (2013). Person-centered emotional support and gender attributions in computer-mediated communication. *Human Communication Research, 39,* 295-316. doi:10.1111/hcre.12006

Stolte, J. (1987). The Formation of Justice Norms. *American Sociological Review, 52*(6), 774– 784. Retrieved from https://www.jstor.org/stable/pdf/2095834.pdf

Thacker, S., & Griffiths, M. D. (2012). An exploratory study of trolling in online video gaming. *International Journal of Cyber Behavior, Psychology and Learning*, *2*(4), 17–33. DOI: 10.4018/ijcbpl.2012100102

Vandebosch, H., & Van Cleemput, K. (2009). Cyberbullying among youngsters: Profiles of bullies and victims. *New Media & Society*, *11*(8), 1349–1371. DOI: 10.1177/1461444809341263

Vermeulen, L., Abeele, M. Vanden, & Bauwel, S. Van. (in press). A gendered identity debate in digital game culture. *Press Start*, *3*(1), 1–16. Retrieved from http://press-start.gla.ac.uk/index.php/press-start/article/view/40

Walther, J. B. (2012). Interaction through technological lenses: Computer-mediated communication and language. *Journal of Language and Social Psychology*, *31*(4), 397–414. DOI: 10.1177/0261927X12446610

Whitty, M. T., & Carr, A. N. (2006). New rules in the workplace: Applying object-relations theory to explain problem internet and email behaviour in the workplace. *Computers in Human Behavior*, *22*(2), 235–250. DOI: 10.1016/j.chb.2004.06.005

Young, S. D., & Jordan, A. H. (2013). The influence of social networking photos on social norms and sexual health behaviors. *Cyberpsychology, Behavior and Social Networking*, *16*(4), 243–7. DOI: 10.1089/cyber.2012.0080

Chapter 3: For Whom the Gamer Trolls

A Study of Trolling Interactions in the Online Gaming Context

The present study aimed to expand our understanding of trolling interactions by examining 10,025 community-reported trolling incidents in the online game League of Legends to determine what characterizes messages sent by trolls, their teammates, and their opponents. To do this, we used a novel method blending content analysis and topic modelling. Contrary to extant literature, our study of complete trolling interactions found striking similarities between teammates' and trolls' chats, with both displaying the negative traits (e.g., exclusionary language) typically attributed to trolls. Findings also suggest that the transition from victim to perpetrator can occur extremely rapidly. This has important implications for the labelling of actors in trolling interactions, for future studies into the trolling cycle, and for theories of computer-mediated communication.

This chapter is based on the following article:

Cook, C., Conijn, R., Antheunis, M., & Schaafsma, J. (2019). For whom the gamer trolls: A study of trolling interactions in the online gaming context. *Journal of Computer-Mediated Communication, 24*, 293-318. doi: 10.1093/jcmc/zmz014

Introduction

As time progresses, our world is becoming increasingly digitalized. In 2013 alone, there were 145 million people globally who were self-described gamers and who played online games as often as 45 to 107 minutes per day (Digital Strategy Consulting, 2013). When this becomes problematic is when one discovers that not all of these players have good intentions. These mal-intentioned people are often called "trolls," and their behavior "trolling" (Buckels et al., 2014; Cook et al., 2018; Fichman & Sanfilippo, 2014; Thacker & Griffiths, 2012). Recently, academia has begun to take an interest in this online phenomenon, providing definitions (Buckels et al., 2014; Fichman & Sanfilippo, 2014; Thacker & Griffiths, 2012), dissecting early cases of the behavior (Herring et al., 2002; Luzón, 2011), and surveying or interviewing various parties involved in the act, from the trolls themselves (Cook et al., 2018; Thacker & Griffiths, 2012) to the moderators of the online communities in which they operate (Shachaf & Hara, 2010).

As a field of study, trolling is known for its multiplicity. Researchers from multiple disciplines have used myriad methods (e.g., vignette studies, surveys, interviews, case studies), and examined numerous populations (e.g., gamers, bloggers, general Internet users) to understand the phenomenon (e.g., Cheng et al., 2017; Luzón, 2011). Few studies, however, have examined how other people interact with trolls, and the majority of those that have took an indirect approach, either by asking participants what they would do in a trolling situation or asking them to reflect on previous trolling experiences (see Maltby et al., 2015; Thacker & Griffiths, 2012). The few trolling interaction case and corpus studies that exist suggest that what both bystanders and victims choose to say has a major impact on the troll's choices (Hardaker, 2010; Herring et al., 2002), but we still do not even know specifically what characterizes the messages of a troll versus the messages of anyone else in the interaction.

To begin to fill this gap, we looked at actual trolling interactions to see how all of the actors involved behaved in real-life trolling interactions, by examining community-reported trolling incidents. We procured a data set of over 10,000 reported cases of trolling), ranging from assisting the opponent team to using offensive language, from the immensely popular online game League of Legends (Riot Games, 2014). Using this data set, we searched for verbal characteristics of trolling interactions (features) which were inherent in the data and examined these to see whether they matched the features identified by previous researchers. In this way, we compared and contrasted a multidisciplinary literature to victim-reported trolling situations in order to answer the following two questions:

RQ1: Do the features portrayed in trolling literature exist in actual trolling interactions?

RQ2: How are these features distributed among the actors (trolls, victims, and bystanders) in the interaction?

To determine what was and was not considered trolling, we relied on the victim's perspective. If the person was reported, the behavior was perceived as trolling. Since we could not determine intent, we followed O'Sullivan and Flanagin's (2014) flaming definition method, and categorized trolling behavior based on victim perceptions. Because of this and the exploratory nature of the study, our definition of trolling was quite broad and included both verbal and behavioral trolling types, defined by Riot Games (2012) as a negative attitude, offensive language, verbal abuse, and assisting the opposing team. For the purpose of the present study, trolling was thus defined as direct or indirect verbal or behavioral aggression that was reported by a League of Legends player under Riot Games' earliest trolling nomenclature (circa 2012), provided this aggression type had also been previously called trolling in gaming-context trolling literature (see Cook et al., 2018, and Thacker & Griffiths, 2012 for complete lists).

Existing Trolling Research

General Features of Trolling Interactions

As explained earlier, trolling research has taken many forms, crossing disciplines, populations, and methods (see Table 1 for an overview). However, it has focused heavily on the person of the troll instead of trolling as a behavior. As such, even when looking at a wide variety of studies, many of the features present in the literature—personality constructs, motivations, emotions, tactics, and more—highlight only the troll, both personally and as a member of the interaction. The current study looked at the messages of all members of the interaction in a gaming context—the reported troll, the members of their team (teammates, composed of one or more victims and one or more bystanders; typically four actors total), and the members of the opposing team (opponents, composed of bystanders; typically five actors total)—to see whether and how the features identified in the literature manifested in a real-life trolling interaction.

Although there was no typology or system of categorization for trolling interactions, the features apparent in the literature gave us some clues as to how the interactions might play out. For example, one of the major findings across most methodologies and disciplines has been that trolls are high in narcissism (Hardaker, 2010; Suler & Phillips, 1998). They enjoy it when the conversation is centered around them, and they tend to seek attention from others by asking a lot of questions (Hardaker 2010) and derailing the conversation (disruption; Table 1), all while pretending to be a serious member of the discussion at hand (deception; e.g., Coles & West, 2016; Kwak et al., 2015). Through this and other means, they display their low communion by promoting discord in the interaction (Buckels et al., 2014; Fichman & Sanfilippo, 2014). As Table 1 shows, some studies have suggested that trolls also demonstrate high agency, which means that they talk profusely, often overrunning the other

members of the conversation in terms of sheer participation (see Buckels et al., 2014). In this trolling dynamic, the other members of the interaction will resort to several different tactics, including refuting the troll's provocative questions and statements or negotiating with the troll to get the discussion back on track (Herring et al., 2002). Herring et al. (2002) affirmed that victims and bystanders may also engage a conflict buffer by telling the others to ignore or block the troll (Herring et al., 2002). Interestingly, this kind of trolling interaction appears to be frequently fueled by boredom on the troll's part (e.g., Maltby et al., 2015; Shachaf & Hara, 2010). The troll is bored with either the website or the game, and wants to pursue a different experience (see Cook et al., 2018). The trolling interaction thus essentially becomes an argument that spirals out of the control of the original actors and into the hands of the troll.

The attention, however, is not the only thing trolls enjoy in these interactions. Researchers have also described trolls as being both psychopathic and sadistic in nature (Craker & March, 2016; March et al., 2017). Table 1 presents many of the tools trolls make use of to elicit pain, including offensive language—often consisting of profanity (Fichman & Sanfilippo, 2014), racism, and sexism (Thacker & Griffiths, 2012)—as well as trash-talking their conversational partners (Cook et al., 2018) and acting generally aggressive and hostile toward victims while instigating the same responses in bystanders (Hardaker, 2010). When this verbal destruction (see Buckels et al., 2014) is enacted by the troll, the interactions are less likely to resemble an argument, and become more of a rant. In this case, the other members of the interaction will likely take more heated action, either by reporting the offender to website or game administrators, or taking revenge and trolling them back, either verbally or behaviorally (Cook et al., 2018; Herring et al.,

Table 1.

Trolling features extracted from extant literature

Category	Feature	Description	Type*
Personality Traits	Sadism	Trolls enjoy the emotional pain they inflict upon their victims.	3
	Narcissism	Trolls tend toward seeking attention; they desire the spotlight in the conversation.	3
	Psychopathy	Total lack of remorse	3
	Impulsivity	Trolls have a lack of self-control.	3
Motivations to troll	Boredom	This is the primary motivation for trolls to troll.	3
	Revenge	Trolling is often provoked by earlier trolling in the conversation.	3
	Emotions	Trolling interactions are heavily-laden with negative emotions, while trolls occasionally display positivity as a function of their sadism	2
Verbal characteristics of trolling interactions	Low communion	Trolls do not value or promote harmony within a group interaction.	2
	High agency	Trolls are highly active within the interaction, talking regularly.	1
	Deception	Trolls will frequently lie to obtain a desired outcome, typically a strong reaction from their victim.	2
	Offensive language	Use of profanity, particularly if in excess.	3
	Aggression/Hostility	Flame-like statements designed to infuriate the other party.	3
	Questions	Frequent use of rhetorical questions.	3
	Repetition	Trolls repeat what they say and do regularly.	1
	Sexism	Any mention of gender in a derogatory sense.	3
	Racism	Any mention of race in a derogatory sense.	3
	Trash-talking	Use of personal insults.	3
Trolling results in the interaction	Destruction	Trolls typically aim to destroy something, be it a reputation, or a group's desired outcome for an interaction.	3
	Disruption	Trolls want to be in control of the conversation, switching the topic to their own ideas even if existing conversation is happening.	3
Victim and bystander tactics	Refutation	Frequent use of comments that are designed to call into question the legitimacy of the previous statement.	3
	Reporting	Calls for reporting someone to an administrator.	3
	Conflict buffers	Calling out trolling behaviour by telling others to stop 'feeding the troll,' or similar statements.	3
	Negotiation	There is discussion in which other actors in the interaction try to reason with the troll.	3

Note. *= This refers to the classification of these features for the purpose of our analyses: 1 = Quantitative, 2 = Complex, 3 = Simple.

2002). Combine this with some of trolls' other tendencies, listed in Table 1, such as the repetition of words and actions (Shachaf & Hara, 2010) and their high levels of impulsivity (Craker & March, 2016; March et al., 2017), and the interaction is likely to escalate to a fever pitch, deeply affecting the emotions of the group (Cheng et al., 2017).

Anonymity and Aggression as Features of Trolling Interactions

Nonetheless, despite the extensive list of features presented in Table 1, there remain two key features missing: anonymity and aggression. Although not a requirement of trolling in its most modern forms (see Cook et al., 2018), due to trolling's fundamentally online nature, anonymity is typically a major factor in its execution. Aggression, or at the very least hostility, is also a necessary component for all of the trolling types listed in our data set. Both of these features would, according to their respective theoretical foundations, escalate trolling situations and cause a strong response from victims. Take, for example, one of the many theories of anonymity's impact on communication: The Social Identity Model of Deindividuation Effects (SIDE) theory (Postmes et al., 1998). According to SIDE, in an anonymous context, people tend to polarize their opinions and expressions to match those of the group. Given the anonymous context of online games, identification theories would predict that trolling interactions would contain indications of argumentation and the polarization of opinions.

Translated to this concrete interaction level, we expected to see team members rally around either the victim or the troll in our data set, therefore producing those argumentation and polarization markers (Postmes et al., 1998). This argumentation between team members and trolls or team members and victims would also be predicted by classic theories of aggression, such as Tedeschi, Smith, and Brown's (1974) theory of coercive action, which states that aggression is the result of a person exercising their coercive power over another

person via threats that result in punishment or a desired behavior. According to this theory, most cases of verbal harassment can be categorized as either the "noxious stimulation" punishment type, meaning the perpetrator introduces a negative stimulus to the victim, or the "social punishment" type, meaning the perpetrator makes the victim look stupid or incompetent in front of their social circle (Tedeschi et al., 1974). In other words, a troll would coerce a player into a desired behavior—an overreaction on said player's part (see Cook et al., 2018, for a complete discussion of trolling goals)—via the application of punishment, which could consist of flaming their victim (insulting them personally as gamers). In addition to these behavioral classifications, the theory also recognizes catalysts or motivations to troll. According to Tedeschi et al. (1974), achieving goals, such as relieving boredom, and being in a negative state of mind (often called being "on tilt" by gamers, see Cook et al., 2018) are both common causes of aggressive behavior that have also been identified in trolling literature (see Buckels et al., 2014; Cook et al., 2018). Nevertheless, the theory also recognizes motivations for the victim to either respond in kind, in "[defense] against the intrusion of others" (Tedeschi et al., 1974, p. 549), bringing a state of equity back to the interaction, or to react with retaliatory norms (i.e., an eye for an eye). Essentially, both theories predict the same thing: an antecedent or trigger is present for the troll from either the current or the previous game(s), the troll exercises coercive power to obtain compliance from their victim and eliminate the unpleasant antecedent, and then the victim retaliates with further aggression.

Trolling also falls under an overarching phenomenon called the online disinhibition effect (e.g., Suler, 2004, 2005). This refers to the discrepancy between a person's online and offline rates of self-disclosure and hostility, and can go one of two ways (Suler, 2004). If people disclose more of themselves or are unusually kind online, this is benign disinhibition, but if the person acts deceptively or in a hostile or unusually aggressive fashion online, this is

toxic disinhibition (Suler, 2004, 2005). Most researchers expect trolls to fall into the toxic disinhibition category (see Cheng et al., 2017; Thacker & Griffiths, 2012), but the categorization of the other members' behaviors is not so clear. Reduced verbal cues (Casale et al., 2015) can make it more difficult for other actors to determine the tone of a troll's messages (i.e., sarcastic or genuine), which would theoretically explain initial attempts by bystanders and victims to negotiate and refute statements rationally (see Herring et al., 2002) before becoming frustrated or angry (see Thacker & Griffiths, 2012). Within the context of Tedeschi et al.'s (1974) theory of coercive action and power, this choice between rationalized discussion and the immediate use of coercive power (victim retaliation) is determined by the victim's self-perception and their perception of the perpetrator. If the victim believes that a discussion will be successful in obtaining their goal—stopping the perpetrator's verbal assault—they may resort to what Herring et al. (2002) labeled refutation and negotiation (see Table 1). If they instead respond in kind, the victim may have determined that either they lack the resources to argue or that the perpetrator is too aggressive to respond to reason, so that retaliating with verbal force is used to reestablish equity in the interaction. Given trolls' tendency to repeat themselves (Shachaf & Hara, 2010) and fully integrate themselves as the key player in the social interaction (Herring et al., 2002), the theory of coercive power would predict trolling interactions ending in victim retaliation. Once more, both theories would put both perpetrator and victim into the toxic category.

In order to both determine the importance of these theories for trolling, while also examining the aforementioned features, we examined which features from literature appeared in our data set, and which actors displayed which features in their messages. Then, we were able to look more globally at the characteristics of each actor and see how closely they matched these theoretical predictions. For example, by the careful analysis of our trolling interaction corpus, we could determine whether victims and bystanders' messages appeared

to be more toxic or benign in natural conversation, thus approximating the prevalence of these two options in the online gaming sphere. We also were able to check for the conversational markers denoting argumentation and polarization of opinion—traditionally operationalized in trolling literature as a combination of "low communion" and "refutation" (see Herring et al., 2002)—that would suggest whether or not SIDE's or Tedeschi et al.'s (1974) theory of coercive action was relevant to trolling research. The term "trolling" may be new, but social interactions and the Internet are not. It is by delving into these established fields that we can begin to knit trolling and theory together and take steps beyond our descriptive foundations into studies examining causal mechanisms and even possible interventions.

Method

The Data Set

Data—consisting of game statistics, chat logs, and report data from reported cases of trolling in the game League of Legends—were obtained from the database administrator of https://tribunal.gc, who had collected it during his tenure serving on League of Legends' Tribunal: a team of high-level players who read user-reports of other players and determined whether the guilty party deserved punishment or not. This system was designed to keep the majority of false reports away from the possibility of punishment. The tribunal itself was introduced in 2011 and closed in 2014 (Riot Games, 2017); all games in this data set took place on the Europe West server during that period. In addition, it should be noted that although there was no informed consent, no individual players could be identified—directly or indirectly—from this set, as the data were completely anonymized.

The data itself took the form of tribunal cases, which consisted of game and chat data from games in which trolling behavior was reported to the Tribunal. These cases specified the

most recent game in which a player was reported for poor behavior, as specified by the reporter, but also listed how many times the player had been reported prior to that particular incident. The types of trolling present in our data set are as follows: assisting the opposing team (1,025 cases), inappropriate username (30 cases), negative attitude (2,931 cases), offensive language (2,562 cases), spamming (167 cases), and verbal abuse (3,343 cases), all of which have been found to be trolling behaviors in the gaming context in extant literature (Cheng et al., 2016; Chesney et al., 2009; Cook et al., 2018; Coyne et al., 2009; Thacker & Griffiths, 2012). Of the aforementioned cases, 33 complete chat logs were found to be in a language other than English, and were thus also removed from our analyses. Our data set thus constituted 10,025 games, and only the chat logs were analyzed in the present study. These logs ranged in length from 1 message to 910 messages ($M = 168.74$, $SD = 104.12$). A glimpse of this data is presented in Figure 1.

These chat logs can be separated into two parts: the chat content and peripheral data. The content consists of all the messages sent by all players throughout the course of the reported game. The complete data set includes 1,697,222 discrete messages sent; these ranged in length from a single emoticon to multiple sentences. The peripheral data consists of all other information: who sent it (troll/teammate/opponent), on which channel it was sent (team chat/opponent chat/all chat), the chosen in-game character of the sender, and the message's timestamp. For the purpose of all analyses, chat messages were divided according to the sender: the troll (troll), one of the four teammates of the troll (teammates), or one of the troll's five opponents (opponents). The troll is the perpetrator, the teammates are composed of one or more victims and one or more bystanders, and the opponents are bystanders.

Figure 1.

Excerpt of the original, uncleaned version of the data.

Case	Time	Champion Name	Association to offender	Channel	Message
1	0:00:21	Udyr	enemy	Team1	gold 2 zed
1	0:00:27	Riven	enemy	Team1	IIII
1	0:00:27	Udyr	enemy	Team1	nice premade lie :o
1	0:00:28	Riven	enemy	Team1	ISI
1	0:00:43	Udyr	enemy	Team1	smiteless pls
1	0:00:57	Udyr	enemy	Team1	smiteless pls
1	0:01:10	Udyr	enemy	Team1	riven?
1	0:01:53	Udyr	enemy	All	report top no help jnh
1	0:08:12	Udyr	enemy	Team1	warded there
1	0:08:17	Riven	enemy	Team1	K
1	0:08:53	Karma	ally	Team2	thx
1	0:09:57	Udyr	enemy	Team1	nice
1	0:11:09	Riven	enemy	Team1	udyr top
1	0:11:48	Udyr	enemy	Team1	riven you know what a ward is about?
1	0:11:59	Riven	enemy	Team1	dnt us see it?
1	0:12:04	Udyr	enemy	Team1	muss weg
1	0:12:09	Udyr	enemy	Team1	mutter stresst
1	0:13:35	Riven	enemy	All	CAMP MORE PLEASE
1	0:14:20	Udyr	offender	Team2	bait
1	0:19:49	Karma	ally	Team2	top no flash
1	0:25:40	Jinx	ally	All	im comming for you riven
1	0:26:05	Riven	enemy	All	u want 1v1 ?
1	0:26:12	Jinx	ally	All	pfft
1	0:28:33	Riven	enemy	All	ok lets continue camping me:)+
1	0:30:44	Jinx	ally	Team2	focus Zed always!
1	0:30:58	Jinx	ally	Team2	and stop feed him!
1	0:31:16	Udyr	offender	Team2	Karma reported
1	0:31:19	Udyr	offender	Team2	Unskilled
1	0:31:25	Udyr	offender	Team2	No ranked for u my friend
1	0:31:31	Janna	ally	Team2	mimimi
1	0:31:40	Karma	ally	Team2	for what ? he has 2 kill in laning phase from me
1	0:31:55	Zilean	enemy	All	lol
1	0:32:06	Udyr	offender	Team2	Didnt i say unskilled player?
1	0:32:19	Karma	ally	Team2	and why you talk about skill you play Udyr
1	0:32:26	Ezreal	ally	Team2	report for unskilled player is useless

Messages sent on the troll's team chat channel could be seen by the troll and their teammates, while only members of the opposing team could see messages sent on their channel. Messages on the global chat channel could be seen by all players. Procedures used to clean the chat data can be found in Appendices 1A and 1B.

Analytical Strategy and Additional Materials

Our first research goal was to determine which of the features present in the literature appeared in our sample of natural dialogues. How we did this depended entirely on the type of feature, as presented in the final column of Table 1. Since our goal was to determine which

features appeared in the data, rather than to impose a set of features on the data, we used three unobtrusive means to explore the data and find our features: two deductive and one inductive.

Deductive Feature Analyses

For repetition and high agency, we were able to use simple statistics and count variables to assess whether the different actors in the interaction (troll, teammates, opponents) repeated themselves or spoke frequently throughout the interaction. We performed these analyses using computer-assisted text analysis software Diction 7.1.3 (Hart, Carroll, & Spiars, 2017) and R 3.3.3 (R Core Team, 2016).

For the rest of our deductive features, we first made the decision to eliminate deception, as to truly determine whether deception was present would have required knowledge of the person's intent, which we did not have. For low communion and emotional valence, we selected two dictionaries that had been previously validated to encompass the various aspects of these variables. For low communion, we chose to employ Diction 7.1.3's commonality dictionary. This dictionary was designed to assess "language highlighting the agreed-upon values of a group and rejecting idiosyncratic modes of engagement" (Digitext, 2017), reflecting the low communion construct (see Buckels et al., 2014) and, to a certain extent, the refutation construct (see Herring et al., 2002) explained in Table 1. The Diction help module, available freely at the software's home webpage, details this and other Diction 7.1.3 dictionaries. The same procedure that was used to examine the low communion feature was also used to examine emotional valence, only with a different dictionary set: the Semantic Orientation Calculator (SO-CAL; Taboada, Brooke, Tofiloski, Voll, & Stede, 2011). This is a two-part, lexicon-based evaluation of semantic orientation, which Taboada et al. (2011, p. 267) defined as "a measure of subjectivity and opinion in text. It usually captures an evaluative factor (positive or negative) and potency or strength (degree to which the word,

phrase, sentence, or document in question is positive or negative." This mirrors Cheng et al.'s (2017) mood construct. For more information, consult Tabaoda and colleagues' 2011 article detailing the tool's construction. SO-CAL is also available for download at https://github.com/sfu-discourse-lab/SO-CAL.

Both of these dictionaries/tools were tested against human coders to minimize the risk of false positives and negatives. Via Amazon's Mechanical Turk (MTurk), 90 participants were paid $5 U.S. dollars each to perform simple coding on 10 cases (each) out of a subset of 300. Participants saw the chat log messages of each case, with speaker letter codes that, unbeknownst to participants, referred to the three actor types (A = troll, B = teammate, and C = opponent); they were asked to rate each speaker on two 5-point scales, from strongly negative (1) to strongly positive (5) and from highly uncooperative (1) to highly cooperative (5). These same 300 cases were then passed through Diction's commonality dictionary and the SO-CAL. When emotional valence scores (SO-CAL and MTurk) were compared, they were found to have only a small correlation ($r = .33$); there was almost no correlation between low communion scores (Diction Commonality Dictionary and MTurk; $r = .01$).

The most likely reason for this discrepancy is the jargon-laced messages both the human coders and lexicons had to parse. Unless the coders were experienced with League of Legends, they would not necessarily have been able to understand the text as it was delivered to the members of the game. This made the text equally difficult for machines to parse, creating a floor effect with the lexicon in many of the cases. Because of this, it is difficult to say whether the MTurk coding was more or less accurate than the machine coding. However, when automatic or human coding is used, using the other type of coding as a test is the best way to validate the initial coding (see de Graaf & van der Vossen, 2015, for an example), and to our knowledge, there are no participant recruitment databases that would have allowed us to include League of Legends experience as a participation criterion. As such, we decided to

exclude the low communion dictionary data, due to the lack of correlation with our human coding data. However, taking the small correlation obtained with emotional valence, its importance in the literature as a predictor of trolling behavior (see Cheng et al., 2017), and the lack of a better alternative into account, we decided to go ahead with our sentiment analyses and interpret our results with caution.

Inductive Feature Analysis

For the other features listed in Table 1, we wanted to be as exploratory as possible. Yet, we still wanted to be able to distinguish between actors and the channels on which they spoke, as our second research question focused on whether the features were displayed by the same actors, as portrayed by extant literature. To do so, we chose to use structured topic modelling, with channel (all, teammates, opponents), role (teammate, troll, opponent), and the interaction between channel and role as covariates. We used spectral initialization and set the model to run for a maximum of 50 expectation maximization algorithm (EM) iterations (for more info, see Roberts et al., 2016). At the outset, we used Lee and Mimno's (2014) algorithm to select the number of topics, but this produced 73 individual topics, which were too many to then analyze manually and name. We therefore tried models that produced 8, 10, or 12 topics, and the 10-topic model seemed to have the clearest interpretation, and was thus kept for analysis.

Table 2.

Linear mixed multi-level models of high agency with role as predictor

Model 1	Estimate (β)	Standard Error	*t*	Conf. Interval (99%)
(Intercept)	17.26	0.12	148.30	(16.95, 17.56)
Actor: Teammate	-5.68	0.12	-45.90	(-5.99, -5.37)
Actor: Opponent	-7.54	0.12	-62.3	(-7.86, -7.22)
Model 2				
(Intercept)	11.58	0.07	157.10	(11.39, 11.76)
Actor: Troll	5.68	0.12	45.90	(5.35, 6.00)
Actor: Opponent	-1.86	0.08	-23.20	(-2.06, -1.66)

Note. Number of observations = 145,662; number of groups = 10,025; ICC = 0.08.

Results

Deductive Analyses

High Agency

Globally, the average amount of messages per case sent by trolls was 30, by teammates was 68, and by opponents was 71. To compare these data statistically, we checked the number of people in the interaction, divided the data to the individual-player level, and sorted by channel (troll's team, opposing team, global chat) and actor (troll, teammate, opponent). Because the opponent actor and opposing team chat categories heavily overlapped—opponents are the only actors who can chat on that channel—actor and channel were entered as predictors into separate linear mixed models. This modelling approach was taken with all multi-level models, due to the uneven nesting of the data. In the first model, we found that—on average—trolls sent 17.26 (*SD* = .12) messages per game, while their teammates sent 11.58 (*SD* = .07) messages per game. Members of the opposing team sent the least, with an average of 9.72 (*SD* = .08) messages per game. Specific results are presented in Tables 2 and 3. To estimate CIs, models were parametrically bootstrapped over 1000 iterations (Bates, Mächler, Bolker, & Walker, 2015).

Table 3.

Linear mixed multi-level models of high agency with channel as predictor

Model 1	Estimate (β)	Standard Error	*t*	Conf. Interval (99%)
(Intercept)	4.99	0.07	72.60	(4.80, 5.17)
Channel: Troll's team	13.62	0.08	163.10	(13.40, 13.84)
Channel: Opposing team	8.96	0.08	105.50	(8.74, 9.17)
Model 2				
(Intercept)	18.61	0.08	238.10	(18.41, 18.80)
Channel: Global chat	-13.62	0.08	-163.10	(-13.84, -13.41)
Channel: Opposing team	-4.66	0.09	-50.60	(-4.91, -4.42)

Note. Number of observations = 145,662; number of groups = 10,025; ICC = 0.10.

Model 2 also showed that the majority of the messages were sent within the team's specific channel (troll's team: $M = 18.61$, SD = .08; opponents' team: $M = 13.95$, SD = .08), as opposed to on the global channel, which everyone could see ($M = 4.99$, $SD = .07$). The effect size on the number of messages sent per actor was quite small ($f^2 = .032$), while channel achieved a medium to large effect ($f^2 = .212$; see Lorah, 2018, for a complete discussion of effect sizes in multilevel models). This confirmed the idea, presented by Buckels et al. (2014), that the average troll sends more messages per game than the average teammate or opponent player. These analyses also affirmed that most of the chats occurred on the troll's team's channel, allowing us to include troll agency in our final visualization.

Repetition

For the full data set containing all available chat messages, 3,561,390 words were identified, 1,798,477 of which were unique (50.5%). However, just as there was one troll and multiple teammates and opponents in the case of high agency, the same problem occurred with repetition: one person will almost certainly use fewer unique words than four or five people combined. As such, instead of using raw percentages to determine the differences between actors, we once again split our data according to the individual-player level for analyses. We found that, on average, trolls said 89 ($SD = 74.99$) words per conversation, 58 ($SD = 40.60$) of which were unique (approximately 73.80%); trolls' teammates said 55 ($SD = 60.88$) words per conversation, 39 ($SD = 35.89$) of which were unique (approximately 81.70%); and trolls' opponents said 43 ($SD = 50.33$) words per conversation, of which 32 ($SD = 31.27$) were unique (approximately 84.55%). The proportions of unique words are significantly different from one another ($F[2,89,815] = 254.00$; $p < .001$; $\eta^2 = .05$), meaning, according to a series of Tukey's honest significant difference tests, that teammates repeated themselves less than opponents, who

repeated themselves less than trolls (all *p*-values < .001; trolls vs. teammates, *d* = .56; trolls vs. opponents, *d* = .80; teammates vs. opponents, *d* = .21). Consistent with Shachaf and Hara's (2010) findings, trolls did indeed have a lower proportion of unique words, compared to their teammates and opponents. We thus chose to include repetition as a variable in our final visualization.

Emotional Valence

A linear mixed model, with actor entered as a predictor of emotional valence, a dimension in which negative scores represent a negative emotional valence and positive scores represent a positive emotional valence, revealed that—on average—trolls' chats registered as more negative (*M* = −.52, *SD* = .02) than their teammates' chats (*M* = −.26, *SD* = .01), which were more negative than the opposing team's (*M* = .00, *SD* = .02) chats, which registered as neutral. Specific results are presented in Tables 4 and 5.

Table 4.

Linear mixed multi-level models of emotion with role as predictor

Model 1	Estimate (β)	Standard Error	*t*	Conf. Interval (99%)
(Intercept)	-0.52	0.01	-39.10	(-0.55, -0.48)
Actor: Teammate	0.26	0.02	14.20	(0.21, 0.30)
Actor: Opponent	0.52	0.02	29.00	(0.47, 0.57)
Model 2				
(Intercept)	-0.26	0.01	-20.6	(-0.29, -0.23)
Actor: Troll	-0.26	0.02	-14.20	(-0.30, -0.21)
Actor: Opponent	0.26	0.02	15.00	(0.22, 0.31)

Note. Number of observations = 53,445; number of groups = 10,025; ICC = 0.02.

Another model with channel entered as a predictor of emotional valence, Model 2 in Table 4, showed that messages sent on the troll's team channel (*M* = −.89, *SD* = .02) were the most negative, followed by those sent on the opposing team's channel (*M* = −.57, *SD* = .02), although they both registered as generally negative. Both were more negative than messages sent on the

Table 5.

Linear mixed multi-level models of emotion with channel as predictor

Model 1*	Estimate (β)	Standard Error	*t*	Conf. Interval (99%)
(Intercept)	-0.19	0.01	-22.40	(-0.21, -0.17)
Channel: Opposing team	-0.38	0.02	-19.40	(-0.43, -0.33)
Model 2				
(Intercept)	0.27	0.01	26.70	(0.25, 0.30)
Channel: Global chat	-1.16	0.02	-75.10	(-1.20, -1.12)
Channel: Opposing team	-0.84	0.02	-43.30	(-0.89, -0.79)

Note. Number of observations = 53,445; number of groups = 10,025. Model 1 ICC = 0.02; Model 2 ICC = 0.04. * = fixed-effect model matrix was rank deficient, so global chat column was dropped.

global channel ($M = .27$, $SD = .01$). The effect size for the actor was very small ($f^2 = .019$), but a small to medium-sized effect was observed for channel ($f^2 = .117$). These results would suggest that, although trolls presented the most emotional content of the three actors in their chats, in reality, this effect was encompassed by the fact that the majority of this negative emotion was expressed on the troll's team's chats. Given their apparent importance for the troll's team, negative emotions were also included in the final visualization.

Inductive Analysis

The results of the structured topic modelling are presented in Table 6. As is evident from the word lists, machine-generated topics are not always easily interpretable by humans. It is thus tentatively that we attempted to map these topics onto the existing extant features detailed in Table 1. In the cases where there appeared to be no connection to existing features, new topic names were given.

Despite aforementioned interpretive difficulties, there were some features we could recognize here. Topic 2, for example, appears to map at least partially onto conflict buffering and refutation (see Table 1), as "play" and "win" suggest someone trying to get people to focus on the game, and "idiot," "farm," and "shut [up]" appear to be refutations of a troll's assertions.

Table 6.

Topics produced from structural topic modelling (STM)

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
Highest Probability	Lol Team Fuck Well_play Baron Drake Jungle	Play Win Farm Sadface Idiot Kill_steal Shut	Nice Troll Yeah Attack_Speed Fck Wait Stupid	Noob Kill Ult Flame Suck Retard Talk	Happyface Bot Wtf Push Stop Lane Dat	OMG Feed Game Lose Care Focus Guy	Good_Game Top Mid Pron Bad Blue Gank	Charname Ty Cait* Ulti Time Tank Blitz*	GJ Ward No_Problem Easy Back Red Ur	Report Switching_Side Lee* Away_from_ Keyboard It Easy Kha*
FREX	Respawn Purple Dance Boost Doge DC Eve*	Charnamet tu Taa Ovo Charnamet un Malph* Mita Farm	Che** Cazzo** Sono** Arrete** Ci** Fois** Dai	Pathetic Heca* Surend Fag Cunt Noob Dildo	Happyface GP Bot Ali* Tris* Recomcecar Arrow	Trynda* Kassa* Ofc Fidle* Solo Lase* Fiddle*	Trynd* Nah Good_Game Funny Jarvan* Trist* Riot	Blitz* Ty Tank Ping Item Sec Follow	Gj Ward Def Back Out_of_Mana Dra* Tribush	Nao Niet** Ere Een** You Anda Echt**

FREX = Words weighted by their overall frequency and how exclusive they are to the topic; * = reference to a character in-game; ** = known word in a language other than English.

Topic 4 is the clearest, with words like “cunt,” “dildo,” and “retard” representing the sexism and offensive language features identified in extant literature. Another striking result from this test is the sheer extent to which game-specific language was used by players. Several shortened character names appeared, as well as many verbs specific to the multiplayer online battle arena (MOBA) genre, such as “gank,” “ult,” “feed,” and “farm.” Topic 1 appears to refer to the jungle region, which is the space between the lanes on the standard League of Legends map, as Baron and Drake are both creatures that appeared in the jungle and Evelyn (Eve) is a character played in the jungle. Topic 7 appears to refer to the “solo-lane,” as the champions mentioned (ex., Tryndamere, Jarvan IV) were typically found in the top or middle lane for most of the game, and both “mid” and “top” were high-probability words. “Blue” refers here to a common, temporary upgrade the middle-lane champion required to be effective, and both the middle and top lanes regularly requested “ganks”: sneak attacks. Topic 8 seems to refer to the bottom lane, as it was the lane that required the most cooperation (ty = thank you), and both Caitlyn (Cait) and Blitzcrank (Blitz) are characters who were frequently played in the bottom lane during the time this data was recorded.

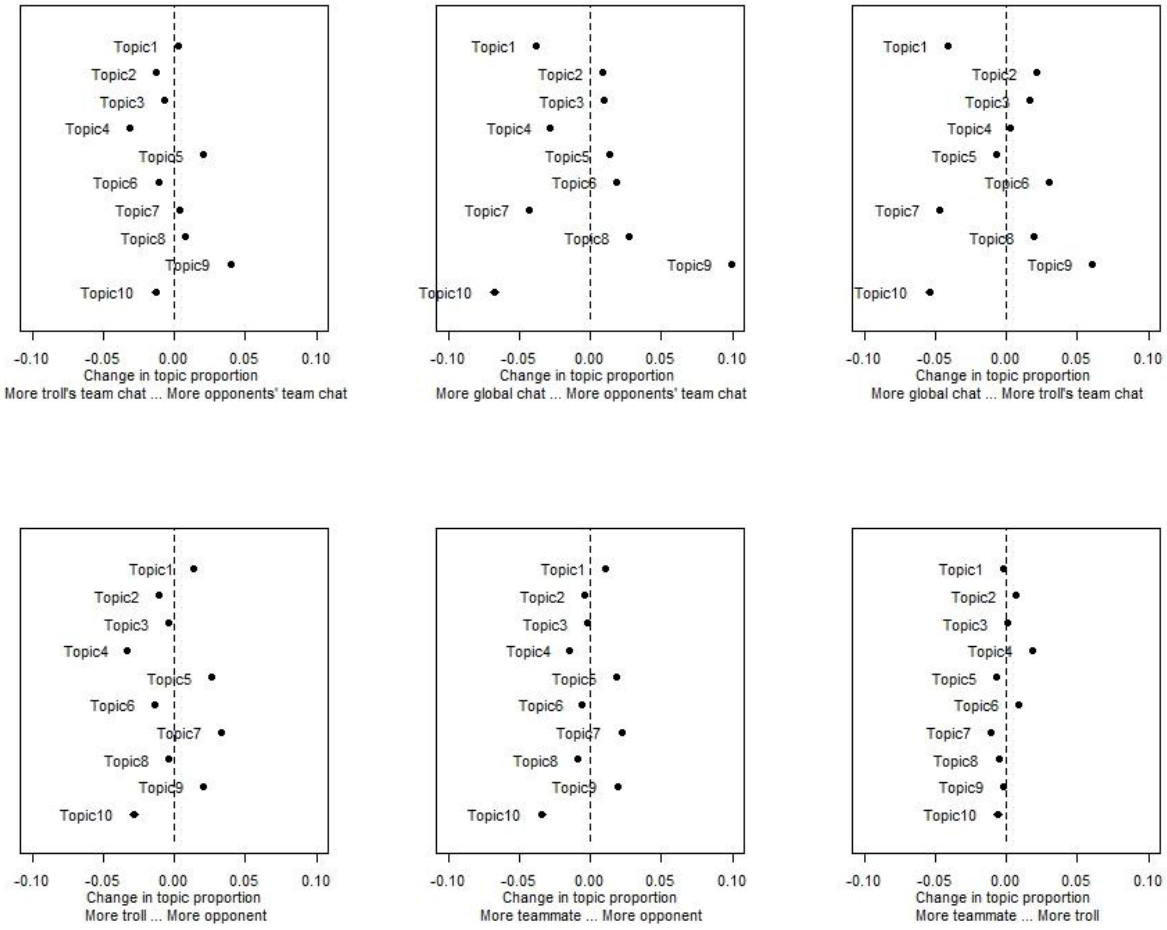
Topics 3, 5, 6, and 9, however, are much less evident. We can see elements of disruption and destruction in Topics 5 and 6, as people demonstrated their exasperation with “OMG” and “WTF,” and we can see the report function appearing in Topic 10, but overall, it is difficult to determine what semantic link the algorithm uncovered. In fact, Topic 3 looks more like a cultural effect than a semantic relationship, given the amount of French and Spanish words that appeared in the list of words mostly exclusive to the topic. This was an unfortunate drawback of using automatic methods. Nevertheless, these words and topics were not used in a vacuum—they represent the features of a trolling interaction—and these interactions happen across both players

and channels. For example, one might assume that Topics 1, 7, and 8 referred to a common in-game practice called “shot-calling,” in which players coordinate their movements across the three lanes and jungle. However, according to the top three graphs in Figure 2, these topics occurred most commonly on the global chat channel, which everyone could see. Thus, it is unlikely that these were shot-calling, as this would mean the players were advertising their positions and plans to the opposing team. At the same time, the topic that seems to reflect the reporting function in League of Legends is also clearly relegated to the global channel, which likely indicates calls for the other team to report a player for trolling that occurred in the other team’s chat channel. By taking the channel into account, we could better distinguish how the features present were probably being used by the various actors.

What the bottom three graphs in Figure 2 demonstrate is essentially the similarity or dissimilarity between the chats of said actors. The more central a topic appears to be in the graph, the more evenly it is distributed through the chat of both actors; the higher the skew, the more specific that topic is to an actor. Irrespective of topics and their names, we can see globally that there were greater distances in topics between trolls and their opponents than between trolls and their teammates. In fact, the graph comparing trolls and their teammates shows that, for all but Topics 4 (offensive language) and 7 (solo-lane shot-calling), the topics were all used equally by both trolls and their teammates, and even these exceptions only deviated slightly. This means that they appeared to talk about the same things, or at least use the same words, frequently. When these two are compared to opponent chats, we can also see that the same topics fell on the side of the troll or their teammates in both graphs. For example, the topic that includes reporting (Topic 10) was almost always used by trolls or their teammates, and rarely by opponents, while opponents seemed to focus more on controlling the map (Topics 1 and 7) and coordinating their

Figure 2.

Results of structural topic modelling (STM).



Note. The top three graphs describe the topical prevalence contrast analyses across channels, while the bottom three graphs describe the same across actors. The features found in each topic are as follows: 1 = Jungle*, 2 = Conflict Buffer & Refutation, 3 = Refutation, 4 = Offensive Language, 5 = Sarcasm, 6 = Anger, 7 = Top lane*, 8 = Bottom lane*, 9 = Teamwork/ Coordination, 10 = Reporting. Starred features are those unique to the MOBA genre.

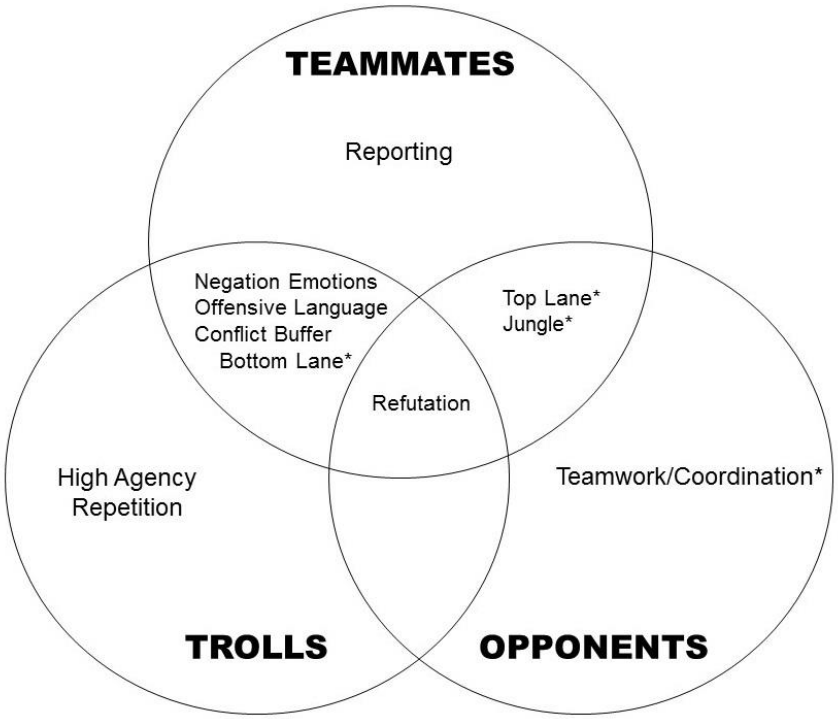
team (Topic 9). They also appeared to be the primary users of sarcasm (Topic 5), although teammates used it more than trolls, suggesting this may be used as a response to trolling. In short, troll and teammate chats appear extremely similar, while opponent chats are distinct.

Summary and Integration of Findings

Based on these findings, we propose the visualization of trolling interactions presented in Figure 3. The one feature unique to trolls is high agency, meaning trolls were typically the most active in the interaction overall. Teammates appeared to be the ones calling for reports, while opponents' chats appeared to be characterized by the coordination of movements across the game map. Across our analyses, these were the only features that—based on their presence on team-specific or global channels and their effect sizes for individual actors—were unique to these members of the interaction.

Figure 3.

Visualization of trolling interactions as they occur in MOBA online gaming.



Note. The three circles describe the three actors in an in-game trolling situation, and the features that occurred in their messages, as well as features that overlapped between actors. Starred items are unique to either the online game medium or the MOBA genre.

Trolls and teammates shared the majority of the features we found. These are also the features that are considered to be characteristic of trolls and trolling in extant literature (Cheng et al., 2017; Herring et al., 2002). Opponents and trolls shared only the element of refutation, which appeared to be shared by all three actors and showed up frequently in the global chat channel. This could signal a general argumentativeness characteristic to in-game chats, but would require the analysis of non-trolling interactions to be certain. In terms of teammates and opponents, only game-specific references to in-game locations and characters were shared. Together, these results illustrate the similarity between perpetrators (trolls) and their victims (usually teammates) in terms of emotional tone and content, as well as the distinct difference between an ongoing trolling situation (trolls' team chat channels) and a normal in-game discussion (opponents' team chat channels).

Discussion

Conclusions

In the present study, our first aim was to determine which features occurred in actual trolling interactions. Overall, we were able to both confirm the presence of many of the features identified in literature, as well as uncover new ones. For example, differences in agency—the number of messages sent per player—were discovered between the actors, indicating that high agency is indeed a sign of a troll. We also found that features often used to distinguish trolls from other actors in the interaction—offensive language and negative mood or attitude—were shared by their teammates, the victims, and the bystanders in the situation. Most prominent among the features, however, were game-based map control features, showcasing just how game-oriented the chats are in this MOBA style of game. The fact that it was primarily the trolls' teammates and opponents making use of these features also suggests that this may be another

form of retort to trolling: a way to exert or regain control over the situation by refocusing attention to game play.

Our second objective was to determine whether the features we found were specific to different actors in the interactions. What we found was that very few features that have been identified in extant literature are actually specific to one actor. The only feature unique to trolls was their high agency. Opponents were distinguished by their high levels of coordination and communion, while teammates had no unique features at all. However, when we examined features that were shared between one or more actors, we could see some important patterns emerging. On the one hand, trolls and teammates exhibited high levels of negativity, leading to a negative team atmosphere. This was further enforced by the finding that the majority of the communication in these trolling incidents took place on the team-specific chat channels, meaning the opposing teams were not even necessarily exposed to the trolling event. Teammates and opponents, on the other hand, seemed to primarily engage with the game via coordination on the map, which could be a ploy to regain control of the situation or simply a way to refocus their attention on the game and avoid the trolling. Overall, these results highlight the influence the troll exerts over the messages sent in their team's chat.

Theoretical and Practical Implications

Another finding was that the trolling interaction features that appear in literature are not always as specific in reality as they are presented in literature. For example, traditionally, negativity and offensive language are treated as troll-specific characteristics (Buckels et al., 2014; Cheng et al., 2017). However, when we performed a basic sentiment analysis on our corpus, we found that there was no significant difference between trolls and their teammates in terms of negativity signals expressed. Our structural topic modelling also showed that trolls did

not appear to use more profanity or offensive language than teammates did in conversation. This means that teammates appeared to be just as offensive and angered in their communication choices as trolls, reinforcing the need to look at the entire trolling interaction when studying the phenomenon, as opposed to just trolls' chats. The few actor-specific characteristics that we found were related to chronemics: repetition and high agency. Neither of these features were mentioned heavily in trolling literature, and yet they were the primary distinguishing marks of a troll in the present corpus. High agency and other such chronemic features may, in fact, be a way forward for the field, as we may be able to more effectively categorize based on how trolls speak, as opposed to what they say.

This lack of actor-defining features led to another, similar conclusion about trolls and their teammates: they are not necessarily easy to distinguish between, based solely on their chat messages. We know from our analyses, including of channels, that the troll's teammates were the most likely victims and bystanders. We also know that they shared negativity and offensive language use. Not only this, but it would appear that trolls also utilized the tactics of their victims and bystanders, as even features typical of a response to trolling (refutation, conflict buffer, reporting) were shared across all actors (see Figure 3). Of course, this similarity of chat was made particularly obvious in light of the fact that opponent chats appeared to be so dramatically different in terms of features, sharing only features common to all actors with trolls, and only two features with teammates (see Figure 2). Altogether, this not only reinforces the power trolls have over the conversation, but also shows that researchers need to be careful when assigning the troll status to a person online, as their victims' and bystanders' chat messages (teammates) seem to strongly resemble their own. As a consequence, researchers are safest when they rely on community testimony regarding trolling, particularly given the fact that trolling differs in its

exact content depending on where it takes place, and given the difficulty in distinguishing trolls from their victims (see Sanfilippo et al., 2017).

Together, these findings about the similarity of trolls' and their teammates' chats also reflect researchers' need to consider the platform carefully when they are designing future trolling studies. Communities are often built around platforms, and communities build shared norms through their interactions on that platform (de Larios & Lang, 2014; Warmelink & Siitonen, 2013). Both trolling researchers and aggression theorists are clear about the impact of community on trolling and hostility: the community is the source of the norms that either encourage or repel trolling behavior (Sanfilippo et al., 2017; Tedeschi et al., 1974). As the present study used data from League of Legends, community norms were bound to reveal themselves in the conversations in our corpus. In fact, we have already seen one such norm: the heavy use of game-specific jargon and abbreviations. The same could be said of other features, such as high agency or offensive language: do they only exist here because it is a League of Legends–specific corpus? It would seem that these are at least partially generalizable, as our list of features was taken from all trolling literature, meaning that many of the features we found here were also present in other contexts, on other platforms (see Table 1). That said, future research should seek to test these, and the other features discussed in Table 1 on different platforms to tease apart community norms from phenomenological characteristics.

Previous research has also suggested that trolling follows a cycle similar to that of cyberbullying, with victims, and occasionally bystanders, becoming perpetrators after repeated exposure to the phenomenon (see Cook et al., 2018; Vandebosch & Van Cleemput, 2009). Our results would suggest that for trolling, one interaction may constitute enough exposure for this transition from victim (teammate) to perpetrator (troll) to occur. Those features that we assume

make a troll's message bad or unpleasant were also present in the other interaction members' messages. Thus, it appears that victims and bystanders who are exposed most directly to the troll's antics start reciprocating within the interaction itself, not waiting for additional exposure before resorting to trolling or using troll-like messages themselves. This has been alluded to in other works discussing trolling interactions (see Hardaker, 2010; Herring et al., 2002), but this is one of the first instances of it being apparent in a data set of this scale. Of course, it should be noted that we cannot confirm causality in this instance, and can only say that the two chat patterns appear highly similar. Longitudinal data would be required to confirm the speed of the trolling cycle with absolute certainty. The present findings do indicate, however, that the cycle for trolling has the potential to progress much quicker than was originally thought.

Our findings also have important implications for other theories of anonymity and aggression. SIDE theory, for example, suggests polarization in conversations and high degrees of negative emotion (Postmes et al., 1998), both of which appeared in our data among trolls and their teammates. This would suggest that verbal trolling—at least the types present in our data set—could indeed constitute an attack on a person's identity, even in the context of a game. Tedeschi et al.'s (1974, p. 551) theory of coercive action also predicts polarization, although it placed the emphasis on trolling being a noxious stimulus that is used to coerce victims into compliance, but ends in "an escalatory cycle of harmful interactions." A similar pattern emerged for the online disinhibition effect. Given the shared negative features of troll and allied chats, it would appear that toxic inhibition is indeed the more popular option in our sample. Since anonymity and aggression both lead to the same outcomes in the case of trolling, it is difficult to determine which mechanism is causal in this instance. It could be, as Sanfilippo et al. (2017) suggested, that this negativity is a feature of the online gaming or League of Legends

communities specifically, and that benign disinhibition may occur more often elsewhere on the Internet. It is also possible that the high state of arousal engendered by being in a game (Lim & Lee, 2009) is at the heart of why trolling interactions seem to escalate so quickly in our corpus, with the retaliatory norm being activated regularly in victims (see Tedeschi et al., 1974). Further research is required to disentangle anonymity and aggression and determine how they interact to produce the kinds of victim responses present in this corpus.

A final practical implication of this research is methodological. As discussed earlier, neither unspecialized human coding nor automatic coding appeared to be sufficient for the examination of game-based chat data. They are some of the best methods available to researchers today (de Graaf & van der Vossen, 2013; Scharkow, 2013)—human coding for its general accuracy, and automatic coding for its time and cost efficiency—but neither appears consistently able to deal effectively with specialized data sets. This is an obvious problem for game studies, but it also exists with any chat data involving jargon. As the data get bigger, traditional methods become less reliable and valid; as the jargon becomes more convoluted, existing resources for automatic methods become insufficient. Therefore, as it stands, researchers interested in said communities will need to ensure that they have access to specialized human coders if they want to perform valid, large-scale content analyses. To avoid incurring the costs associated with large-scale manual coding, computational scientists and other researchers of various specialties will need to collaborate, and either tailor existing algorithms and data-cleaning protocols to the communities in question or build new protocols that can be adapted to these specific data sets.

Limitations and Future Directions

Although novel and fruitful, this study is not without limitations. The method we used (topic modelling and content analysis) is, though largely automatic and cost-effective, not the only

option. When working with millions of discrete messages in a limited timeframe, however, automatic methods are the logical first step. Nevertheless, as we discovered throughout the data-cleaning process, League of Legends players tended toward using small words, which prevented the usage of more advanced cleaning techniques. This, in turn, could overload topic models with smaller content words and, occasionally, even let non-content words slip through. It also meant that even human coders, unless they had prior knowledge or experience with the game, would have had a difficult time understanding and parsing the chat logs. It is important to note that this may be an artefact of the MOBA genre, which is generally fast-paced, leaving little time for lengthy chats when compared to other gaming genres.

That said, there are also interaction-level analyses that are difficult with automatic means, such as detecting sarcasm in valence analyses. An example of this would be the usage of emoticons. Due to the automatic nature of data processing and cleaning in the present study, we could not examine emoticons individually to determine their exact meanings within the contexts of the conversations. For parsimony's sake, we therefore categorized emoticons in the data set as either happy or sad, when in fact they could have been expressing more nuanced emotions, such as sarcasm or a sense of mischievousness. Although efficient, this practice limited the amount of emotional information that could be gleaned from the corpus. There is also the possibility of false reports, as with data of this size, it is unrealistic to go through each case manually to confirm that the proposed perpetrator actually committed the alleged offense, and we did not have the automatic means to confirm with absolute certainty whether an offense was committed or not. In short, using automatic content analysis techniques comes with a potential lack of precision that manual coding typically ensures.

There is a possible step forward here which would help advance trolling research tremendously: the creation of platform-specific dictionaries or lexicons to make automatic methods a more accurate option. These dictionaries would theoretically include not only known words, but also abbreviations, emoticons, and even phrases that are specific to either gaming genres, in the case of online games, or perhaps to Reddit or Twitter vernacular, in the case of other social platforms. Doing this would allow processing of the incredible amounts of data that these major companies produce (Marr, 2018), without having to sacrifice quality due to the data's inherent messiness. If human coders could be removed from the process, automatic data processing techniques would be not only efficient and cost-effective, but also reliable and fully validated ways for research teams to gather population-wide trends in trolling and other phenomena involving online communities.

Finally, it is important to note that the present model was based on game data—even more specifically, MOBA data—and is thus currently valid only for game-related trolling. That said, many of the features that were included in the model can be applied outside of gaming, particularly since they were initially discovered in other data, such as forums (e.g., Herring et al., 2002), or via non-gamer populations (e.g., Maltby et al., 2015). Thus, this model should be tested in other media to determine its applicability elsewhere. This would also open the discipline to cross-platform studies, as currently, most studies have focused on one particular medium: either games (Cook et al., 2018; Thacker & Griffiths, 2012), public websites (Herring et al., 2002; McCosker, 2014), or social networking sites (Craker & March, 2016; March et al., 2017). By opening trolling up and examining its various components—behavioral types, actors, and platforms—we can deepen our understanding of this puzzling and timely online phenomenon.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1-48. DOI: 10.18637/jss.v067.i01

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

Casale, S., Fiovaranti, G., & Caplan, S. (2015). Online disinhibition: Precursors and outcomes. *Journal of Media Psychology*, *27*, 170-177. DOI: 10.1027/1864-1105/a000136

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017, February). *Anyone can become a troll: Causes of trolling behavior in online discussions*. Paper presented at the 20th conference on Computer-supported cooperative work and social computing, Portland, Oregon.

Cheng, J., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2016, May). *Antisocial behavior in online discussion communities*. Paper presented at the 9th annual conference on web and social media, Oxford, United Kingdom.

Chesney, T., Coyne, I., Logan, B., & Madden, N. (2009). Griefing in virtual worlds: Causes, casualties and coping strategies. *Info Systems Journal*, *19*, 525-548. DOI: 10.1111/j.1365-2575.2009.00330.x

Coles, B. A., & West, M. (2016). Trolling the trolls: Online forum users constructions of the nature and properties of trolling. *Computers in Human Behavior*, *60*, 233-244. DOI: 10.1016/j.chb.2016.02.070

Cook, C., Schaafsma, J., & Antheunis, M. L. (2018). Under the bridge: An in-depth examination of online trolling in the gaming context. *New Media and Society*, *20*, 3323-3340. DOI: 10.1177/1461444817748578

Coyne, I., Chesney, T., Logan, B., & Madden, N. (2009). Griefing in a virtual community: An exploratory survey of Second Life residents. *Journal of Psychology, 217*, 214-221. doi: 10.1027/0044-3409.217.4.214

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. DOI: 10.1016/j.paid.2016.06.043

Digital Strategy Consulting (2013). Global online gaming trends: Asia Pacific and Europe lead the way. Retrieved from http://www.digitalstrategyconsulting.com/intelligence/2013/06/global_online_gaming_trends_asia_pacific_and_europe_lead_the_way_infographic.php (accessed 7 July, 2017).

Digitext (2017). Diction overview. Retrieved from http://www.dictionsoftware.com/diction-overview/ (accessed 10 July, 2017).

Fichman, P., & Sanfilippo, M. R. (2015). The bad boys and girls of cyberspace: How gender and context impact perception of and reaction to trolling. *Social Science Computer Review*, *33*, 163-180. DOI: 10.1177/0894439314533169

de Graaf, R., & van der Vossen, R. (2013). Bits versus brains in content analysis. Comparing the advantages and disadvantages of manual and automated methods for content analysis. *Communications*, *38*, 433-443. DOI: 10.1515/commun-2013-0025

Hardaker, C. (2010). Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research, 6*, 215-242. DOI: 10.1515/JPLR.2010.011

Hart, R. P., Carroll, C., & Spiars, S. (2017). Diction (version 7.1.3) [Computer software]. Retrieved from https://www.dictionsoftware.com

Herring, S., Job-Sluder, K., Scheckler, R., & Barab, S. (2002). Searching for safety online: Managing "trolling" in a feminist forum. *The Information Society*, *18*, 371-384. DOI: 10.1080/01972240290108186

Kwak, H., Blackburn, J., & Han, S. (2015, April). *Exploring cyberbullying and other toxic behavior in team competition online games*. Paper presented at the 33rd annual conference on human factors in computing systems, Seoul, South Korea.

de Larios, M., & Lang, J.T. (2014). Pluralistic ignorance in virtually assembled peers: The case of World of Warcraft. *Games and Culture*, *9*, 102-121. DOI: 10.1177/1555412013512894

Lim, S., & Lee, J.-E.R. (2009). When playing together feels different: Effects of task types and social contexts on physiological arousal in multiplayer online gaming contexts. *CyberPsychology & Behavior*, *12*, 59-61. DOI: 10.1089/cpb.2008.0054

Lorah, J. (2018). Effect size measures for multilevel models: Definition, interpretation, and TIMSS example. *Large-scale Assessments in Education*, *6*, 8. DOI: 10.1186/s40536-018-0061-2

Luzón, M. J. (2011). 'Interesting post, but I disagree': Social presence and antisocial behaviour in academic weblogs. *Applied Linguistics*, *32*, 517-540. DOI: 10.1093/applin/amr021

Maltby, J., et al. (2015). Implicit theories of online trolling: Evidence that attention-seeking conceptions are associated with increased psychological resilience. *British Journal of Psychology*, *107*, 448-466. DOI: 10.1111/bjop.12154

March, E., Grieve, R., Marrington, J., & Jonason, P. K. (2017). Trolling on Tinder® (and other dating apps): Examining the role of the dark tetrad and impulsivity. *Personality and Individual Differences*, *110*, 139-143. DOI: 10.1016/j.paid.2017.01.025

Marr, B. (2018, May 21). How much data do we create every day? The mind-blowing stats everyone should read. *Forbes*. Retrieved from https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#14efe5d660ba

McCosker, A. (2014). Trolling as provocation: YouTube's agonistic publics. *Convergence: The International Journal of Research into New Media Technologies, 20,* 201-217. DOI: 10.1177/1354856513501413

Postmes, T., Spears, R., & Lea, M. (1998). Breaching or building social boundaries? SIDE-effects of computer-mediated communication. *Communication Research, 25,* 689-715. doi: 10.1177/009365098025006006

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.

Riot Games (2017). League of legends news: Player behavior. Retrieved from http://na.leagueoflegends.com/en/news/game-updates/player-behavior (accessed 7 July, 2017).

Riot Games (2014). Player numbers. Retrieved from https://www.riotgames.com/tags/player-numbers (accessed 7 July, 2017).

Riot Games (2012). How does the Tribunal work? Retrieved from http://forums.na.leagueoflegends.com/board/showthread.php?t=2136334 (accessed 16 October, 2018)

Sanfilippo, M., Yang, S., & Fichman, P. (2017). Trolling here, there, and everywhere: Perceptions of trolling behaviors in context. *Journal of the Association for Information Science and Technology, 68,* 2313-2327. DOI: 10.1002/asi.23902

Scharkow, M. (2013). Thematic content analysis using supervised machine learning: An empirical evaluation using German online news. *Quality & Quantity, 47*, 761-773. DOI: 10.1007/s11135-011-9545-7

Shachaf, P., & Hara, N. (2010). Beyond vandalism: Wikipedia trolls. *Journal of Information Science, 36*, 357-370. DOI: 10.1177/0165551510365390

Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior, 7*, 321-326. doi: 10.1089/1094931041291295

Suler, J. (2005). The online disinhibition effect. *International Journal of Applied Psychoanalytic Studies, 2*, 184-188. DOI: 10.1002/aps.42

Suler, J. R., & Phillips, W. L. (1998). The bad boys of cyberspace: Deviant behavior in a multimedia chat community. *CyberPsychology & Behavior*, *1*, 275-294. DOI: 10.1089/cpb.1998.1.275

Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, *37*, 267-307. Retrieved from http://www.sfu.ca/~mtaboada/nserc-project.html

Tedeschi, J.T., Smith, R.B., & Brown, R.C. (1974). A reinterpretation of research on aggression. *Psychological Bulletin, 81*, 540-562. DOI: 10.1037/h0037028

Thacker, S., & Griffiths, M. D. (2012). An exploratory study of trolling in online video gaming. *International Journal of Cyber Behavior, Psychology and Learning*, *2*, 17-33. DOI: 10.4018/ijcbpl.2012100102

Warmelink, H., & Siitonen, M. (2013). A decade of research into player communities in online games. *Journal of Gaming & Virtual Worlds, 5*, 271-293. DOI: 10.1386/jgvw.5.3.271_1

Chapter 4: Trolls Without Borders

A Cross-Cultural Examination of Victim Reactions to Verbal and Silent Aggression Online

Trolling – the online exploitation of website, chat, or game mechanics at another user's expense – can and does take place all over cyberspace. It can take myriad forms, as well – some more overt, like trash-talking an opponent in a game, and some more covert, like misdirecting a new forum user to break a rule. However, despite this variety, there are few to no studies comparing the effects of these differing trolling types on victims. In addition, no study has yet taken into account users' offline cultural context and norms into the victim experience. To fill this gap in the literature, the present study put participants from three culturally-distinct countries – Pakistan, Taiwan, and the Netherlands – in a simulated trolling interaction using Williams and colleagues' (2000) Cyberball game. Participants were either flamed (read: harshly insulted) or ostracized by a member of their own cultural group (ingroup) or a minority member (outgroup), and the participants' emotional responses, behavioural intentions toward the other players, and messages sent during the game were taken as indicators of their response to the trolling. Results showed that our Taiwanese sample used the most reactive aggression when trolled and our Dutch sample was the most passive, in line with Anjum and colleagues' (2019) results. In addition, ostracism generally produced the desire to repair relationships, irrespective of cultural context, and perpetrator culture (ingroup or outgroup) only produced an effect in the behavioural intentions of our Pakistani sample. Overall, it would appear that online and offline culture interact to produce the variety of responses to trolling seen in extant literature. Additional implications for future research into computer-mediated communication and online aggression are also discussed.

This chapter is currently under review (second round) as:

Cook, C., Schaafsma, J., Antheunis, M.L., Shahid, S., Lin, T.J.-H., & Nijtmans, H. (2020). Trolls without borders: A cross-cultural examination of victim reactions to verbal and silent aggression online. *Frontiers in Psychology – Cultural Psychology*. Manuscript under review.

Introduction

Around the world, anyone who uses a social network, comments on YouTube, or plays an online game even casually is at risk of experiencing online hostility, referred to as trolling (e.g., Buckels et al., 2014). In the online gaming world, experiencing trolling is a kind of rite of passage (Cook et al., 2018), and trolling behaviour can take a myriad of forms. For example, 'trash talking' is a commonly used technique whereby a player insults or abuses another player, often with the intent to annoy him or her or to derail the game (e.g., Cook et al., 2018). In other instances, the trolling behaviour is much more subtle, such as a player ignoring a teammate's cries for assistance, or purposely lengthening the game by refusing to take their turn, known as 'bad manner' in gaming communities (Arjoranta & Siitonen, 2018). In such cases, the intent of the aggressor is often either hidden, or at least more ambiguous.

How do people react to these different forms of online aggression? At first glance, it may appear that especially overt forms of aggression such as insults – called flaming in the online context (O'Sullivan & Flanagin, 2003) – should be particularly aversive, as they form a direct threat to people's self-esteem and reputation, to which they typically respond with embarrassment or anger (e.g., Liu et al., 2018). One could also make the case, however, that more covert, non-verbal forms of aggression such as ostracism should be equally or perhaps even more aversive, as they threaten people's fundamental needs such as their sense of belonging and their self-esteem, but also their sense of existence and recognition (e.g., Williams, 2009). Whereas people do get some attention and recognition, albeit negative, when they are being insulted, ostracism sends the message that they are unworthy of attention at all (see Filipkowski & Smyth, 2012; Hartgerink et al., 2015; James, 1950). When people experience this what researchers often call a 'social death' (e.g., Williams, 2007), they have been shown to respond in

a variety of ways. Although there is some evidence that they may, under certain circumstances, try to seek re-inclusion (e.g., Ouwerkerk et al., 2005), various studies also show that they can respond with anger and aggression (e.g., Hales & Williams, 2018) or even seek solitude (Leitner et al., 2014).

To our knowledge, however, there has been no or little research comparing people's reactions to verbal and non-verbal forms of online aggression and so, at present, it is not clear whether they result in similar or different responses. The main goal of the present study is to address this issue, by examining how people from different cultural contexts respond to being flamed or being ostracized by in-group or out-group members. We use a cross-cultural angle and rely on samples from differing contexts because cultures have different norms when it comes to responding to threats to the (social) self, and also have different norms about the need to maintain harmony and fit into the group, which could impact how people react to insults and ostracism (e.g., Bond et al., 1985). For instance, according to much of Cohen's and Nisbett's work (e.g., Cohen & Nisbett, 1997), in cultures where honour is more salient and where maintaining respect is a central virtue, people should respond quickly and even aggressively when their reputation is threatened with witnesses present, particularly when the perpetrator is not a member of their social group (Anjum et al., 2019; Allpress et al., 2014; Giner-Sorolla, 2019). Yet, in cultural settings where maintaining face or avoiding face loss and ingroup harmony are more important, people may be more likely to feel embarrassed by a flame and prefer to avoid confrontation, especially if the aggressor is an ingroup member (e.g., Lee, Leung, & Kim, 2014).

To examine how people across different cultural contexts respond when flamed or ostracized, we conducted an experiment in Pakistan – which is generally considered an honour

culture (see Anjum et al., 2019) – and Taiwan, which has been described as a face culture (see Ting-Toomey et al., 1991). As an additional comparison group, we also included participants from the Netherlands, where concerns about face and honour are likely to be less salient or prevalent, and where people theoretically develop a sense of self that is relatively insensitive to the influence of others (see Markus & Kitayama, 1991; Leung & Cohen, 2011). We were principally interested in how angry and how embarrassed or humiliated participants across these different settings would feel following insults or ostracism by ingroup or outgroup members, and whether they would be motivated to retaliate or would prefer to withdraw or to restore relationships instead.

Theoretical Background

Honour Concerns and Reactions to Verbal and Silent Aggression

As mentioned in the introduction, the present literature on how people respond to threats to the self would suggest that people from a culture in which honour is salient should be particularly sensitive to overt, verbal forms or aggression such as flaming. Of particular importance in this regard is Leung and Cohen's (2011) theory regarding how different cultures conceive of reputation as a concept. Although there are three conceptions according to this theory – dignity, honour, and face – honour is arguably the most researched in terms of its connection to aggression (see Nisbett & Cohen, 1996). Leung and Cohen (2011) describe honour as being a combination of how people see themselves and how society sees them: "honour must be claimed, and honour must be paid by others" (Leung & Cohen, 2011, p. 509). In other words, it is up to each person to both develop their own reputation (honour), and also to treat other people with the respect their honour deserves. When someone does not pay a person respect according to their honour, the victim loses their honour, and has to fight or punish the offender to

regain it. This tendency to employ ‘reactive aggression’ (Ang et al., 2014) following threats to one’s honour has been shown with relative consistency in empirical work among cultures that conceive of reputation in this way. Cohen and Nisbett (1994), for example, found that men in the southern United States endorse violence when they are trying to defend their honour or the honour of their family, while Uskul and Cross (2018) found repeatedly that their Turkish participants were particularly likely to retaliate aggressively when they perceived a loss of honour via insult or accusation. These findings provide support for the idea that in a setting where honour is valued, the cultural norms are more likely to dictate that retaliation to regain honour is justified (e.g., Glick et al., 2016).

Yet, while members of honour-valuing cultures may react with reciprocated aggression to a flame due to the obvious insult to their honour, there is also reason to believe that they may be less likely to defend their honour when ostracized. Although neither aggression option in the present study is pleasant to experience as a victim, there is a sharp distinction between flaming and ostracism when it comes to the idea of insult. Flaming is far more direct, as it consists of verbal insults and hostility directed at the victim, often peppered with profanity and expressed with a liberal use of the caps lock button (O’Sullivan & Flanagin, 2003). When a person is being ostracized, however, the reason behind the ostracism is not expressed, and so the victims are left to their own devices when it comes to interpreting the hostility as insulting or otherwise (Williams, 2009). In short, ostracism is a form of aggression that can and often does hurt (see Williams, 2009; Williams et al., 2000), but it is not necessarily insulting. Empirical work suggests that in such situations when a direct insult is not perceived, people from honour-valuing cultures may actually prefer peaceful solutions to their conflicts (Harinck et al., 2013; Pfundmair et al., 2015). In Harinck and colleagues’ (2013) study, for instance, honour-valuing participants

who were told to imagine being in a conflict, but not insulted, still agreed to work with the other party in the conflict to solve the issue together, contrary to the participants who were insulted. In Pfundmair and colleagues' (2015) study, they also found that members of more collectivistic cultures – which honour-valuing cultures generally are (see Anjum et al., 2019) – do not share the aggressive intentions of dignity-valuing cultures when faced with ostracism.

All that said, theory and empirical work would both suggest that people from honour-valuing cultures respond to ingroup and outgroup members differently (e.g., Cross, Uskul, Gerçek-Swing, Alözkan, & Ataca, 2013). More specifically, there is reason to believe that in honour-valuing cultures, verbal aggression by outgroup members - operationalized in the present study as flaming - should result in anger and a stronger desire to retaliate than if the flaming was performed by an ingroup member. This difference is rooted in the beliefs in honour-valuing cultures that honour needs to be defended when threatened, and that honour is shared amongst ingroup members (Leung & Cohen, 2011). When a person retaliates against an aggressor, particularly when the aggressor is employing such an overt tactic as flaming (see Cook et al., 2018), they are fundamentally risking their relationship with that person. When the perpetrator is an outgroup member, there is no existing relationship to threaten, and so the maxim of defending one's reputation is free to be pursued by the honour-valuing victim (Leung & Cohen, 2011; Severance et al., 2013). However, when the perpetrator is an ingroup member, there is a critical pre-existing relationship that could be threatened by retaliating (Severance et al., 2013; Uskul & Over, 2014). When people aggress a close ingroup member in this kind of cultural context, they are risking their own social standing (Leung & Cohen, 2011; Severance et al., 2013). Thus, the risks associated with retaliating against an ingroup offender are likely to be judged too high, and

so reactive aggression should theoretically be reserved for aggression originating from an outgroup perpetrator.

Silent forms of aggression such as ostracism, however, should produce more embarrassment than anger at having caused an unknown offense leading to being ostracized, as well as a stronger tendency to engage in repairing the relationship when the perpetrator is an ingroup member, as opposed to an outgroup member. Part of this expectation comes from the fact that embarrassment is a negative, self-conscious emotion that is produced when a person's identity is being threatened in some way (Chen et al., 2020; Dasborough et al., 2020). Ostracism is a potential threat to a person's social identity, particularly when coming from an in-group member (e.g., Severance et al., 2013). This is likely to be amplified by honour-valuing cultures' emphasis on the ingroup (typically the family) and interconnectedness (Leung & Cohen, 2011; Markus & Kitayama, 1991; Severance et al., 2013). The ingroup is generally the person's source of reputation, and often of basic necessities like food and shelter (Severance et al., 2013). If a person is being flamed by an ingroup member, the connection to the ingroup is only being risked if they choose to retaliate. However, if people are being ostracized by their ingroup, then they are potentially losing not only their social standing and sense of belonging, but also their livelihood. Ostracism is also often used as a form of punishment in certain communities or populations (see Freedman et al., 2016; Hales et al., 2016; Poon & Chen, 2016), so they may be embarrassed at having done something to deserve this punishment. In such a situation, it is more commendable to ignore the offense to preserve honour and relationships (see Cross et al., 2013 for examples), but repairing the relationship would be even more desirable, as it could lead to the victim's reconnection to their source of security (see Severance et al., 2013). This hope for reconciliation within the in-group has been demonstrated repeatedly when it comes to negative self-conscious

emotions (shame, guilt, and embarrassment), particularly when it comes to ingroup members witnessing other ingroup members transgress (Allpress et al., 2014; Giner-Sorolla, 2019) We can see also this in action in Uskul and Over's (2014) study of farmers and herders, in which farmers – who are more dependent on the family unit for their sustenance – responded less negatively to being ostracized by strangers than herders, who typically depend on the patronage of strangers to survive. In essence, when an honour-valuing culture member's connection to the ingroup is being threatened by ostracism, they will theoretically try to repair that connection; when no such connection exists, as with the outgroup, they have no need to engage in said reparative actions.

However, it should be noted that most studies on honour-valuing cultures and aggression were conducted in face-to-face settings if it is a physical experiment or observational study (e.g., Cohen & Nisbett, 1997), or an imagined face-to-face setting if it is a survey or vignette study (e.g., van Osch et al., 2013). Although we recognize this difference between the present study and extant literature, we continue to base our expectations on what we do know, as very few studies have examined honour-valuing cultures in the online context. For instance, we expect that honour-valuing people will react with more anger to outgroup members than ingroup members that flame them (H1a). The two studies that do focus on honour concerns online - Günsoy and colleagues' (2015), as well as Pearce and Vitak's (2016) – results confirm that honour-valuing people are just as concerned about protecting both their own honour and the honour of their families on- and offline. Thus, even though we have far fewer bystanders in the present study than on social media, the idea of protecting one's honour should not be less salient than if we conducted the study offline. Much of extant literature also presents a different social distance between the ingroup members than the present study. While these studies, along with many offline studies (e.g., Severance et al., 2013), focus on close ingroup members like parents,

the present study aims to simulate an average online trolling experience, which typically involves strangers (see Cook et al., 2018; Synnott et al., 2017). Although we do make a distinction between ingroup and outgroup, our ingroup – fellow students of the same university and nationality – is unlikely to be as important to our participants as their family members. However, again, there are surprisingly few studies that deal with the intersection of cultural values and tie strength, and those that do tend to focus on entrepreneurship and business (e.g., Johnson et al., 2013; Ma et al., 2011), or how social groups harness social media to build new ties and strengthen existing ones (e.g., Gonzales, 2017; Kwak & Kim, 2017; Verdery et al., 2018). Thus, although the literature is admittedly scant in an online context, what exists appears to support our hypotheses.

Face Concerns and Reactions to Verbal and Silent Aggression

Just as existing literature predicts that members of honour cultures will respond differently to flaming and ostracism, it also predicts that members of face-valuing cultures will react to ostracism in much the same way as members of honour-valuing cultures, but will not retaliate when flamed. Unlike the construct of honour, face is exclusively an external evaluation of a person's worth, and thus cannot be gained, but can be easily lost (Leung & Cohen, 2011; Hashimoto & Yamagishi, 2013). This also means that in social interactions, it must be carefully preserved, and unlike honour, it cannot be regained via defence. To prevent the loss of face, researchers have posited that in cultures where face is valued as people's primary form of reputation, they will avoid conflict when possible, ignoring perceived slights and hostility in order to preserve the face of everyone involved (Hashimoto & Yamagishi, 2013). Empirically speaking, this is most evident in cyberbullying research, where students from countries in East Asia – often presented in cross-cultural studies as "face cultures" – will avoid confronting

aggressors directly for fear of losing face. Instead, these students try to seek support from others after the fact, or report the aggressor to an authority figure privately (Li, 2008; Ma & Bellmore, 2016).

The tendency to avoid conflict to preserve face in any aversive situation – flaming or ostracism – has been empirically demonstrated in both positive and negative situations (e.g., Bresnahan et al., 2002; Peng & Tjosvold, 2011), meaning that it is considered equally reprehensible to seek and accept praise without demonstrating humility and to retaliate against an aggressor (Kim & Cohen, 2010; Lee et al., 2014). In essence, the literature suggests that members of face-valuing cultures are driven by a desire to avoid individual attention, positive or negative, in all social situations. This is contrary to honour-valuing cultures, where people would theoretically be more likely to defend their reputation when it is being directly threatened via insult. Whether a person is being praised or insulted in a face-valuing cultural context, they are unlikely to give a strong reaction, instead choosing to withdraw, as being too proud or being too aggressive would result in an irredeemable loss of face (Leung & Cohen, 2011).

It is important to note, however, that most face-valuing cultures share the ingroup-centric values common to most honour-valuing cultures (Anjum et al., 2019; Severance et al., 2013). As such, we expect the experience of being flamed by an ingroup member to be a more intensely negative experience for people from face-valuing cultures than the experience of being flamed by an outgroup member. Nevertheless, when people from a face-valuing culture believe that an ingroup member is flaming them - much like the case with honour-valuing culture members - retaliation may not be an option, as this would risk damaging the relationship with the ingroup (see Leung & Cohen, 2011; Severance et al., 2013). This may be compounded by the avoidance tendency found in many empirical studies on how people from face-valuing cultures respond to

threats and aggression in general (see Hashimoto & Yamagishi, 2016; 2013). Theoretically, therefore, an ingroup perpetrator would put people from face-valuing cultures into a particularly uncomfortable position when being insulted, as their primary goal is to preserve face and fit in with the ingroup (see Severance et al., 2013). This would, according to Kitayama, Mesquita, and Karasawa (2006), lead to what they call negative engaging emotions, such as embarrassment or shame, as well as a desire to withdraw. Although they may also want withdraw from conflict in the face of an outgroup member due to their conflict avoidance (see Hashimoto & Yamagishi, 2016; 2013), the same risk of being disconnected from the ingroup – their source of reputation and sometimes basic necessities (Severance et al., 2013) – does not exist to the same degree in this circumstance.

In the case of ostracism, we expect to see the same pattern with people from face-valuing cultures as we expect with honour-valuing cultures: an increase in attempts to repair the relationship when faced with an ingroup perpetrator when compared to an outgroup perpetrator. As is the theoretical case with flaming, ostracism inherently violates the collectivistic, ingroup-centric values of face-valuing culture members: preserving social harmony through fitting in and avoiding conflict (Peng & Tjosvold, 2011; Pfundmair et al., 2015; Severance et al., 2013). Just as is the theoretical case for people from honour-valuing cultures, being ostracized by an ingroup member adds an extra dimension of rejection for members of face-valuing cultures, as this means that they are being actively separated from their support network and source of reputation (see Severance et al., 2013). As in the case of honour-valuing people, this may not work in exactly the same way online with strangers as it does offline with family members. However, participants are aware that they are being recorded in-game, though personal anonymity is

guaranteed; this should still elicit at least some face concerns, even if they are not as strong as if it was an in-person family-based situation.

Due to their inherent motivation to fit in and preserve harmony, although their preference would be to withdraw and avoid conflict, face-valuing culture members should be more likely to try and repair the relationship when ostracized by an ingroup member, restoring the harmony they allegedly prize (Hashimoto & Yamagishi, 2016). Again, this is the case with most cultures when it comes to ingroup transgressions (e.g., Allpress et al., 2014; Giner-Sorolla, 2019) and self-conscious emotions (Chen et al., 2020; Dasborough et al., 2020); this effect is simply theoretically amplified by the rejection-avoidance inherent to the cultural logic of face (Leung & Cohen, 2011; Hashimoto & Yamagishi, 2013; 2016). When an outgroup member is the perpetrator, like the situation with honour-valuing culture members, there is no pre-existing relationship to repair, and so they are not losing the vital connection to the ingroup (Severance et al., 2013). As such, there is no need to engage in relationship reparation with outgroup members and ignoring the offense and withdrawing should be enough to preserve the face of all parties involved (Leung & Cohen, 2011; Markus & Kitayama, 1991; Severance et al., 2013).

The Present Study

As was mentioned in the introduction, the main goal of this study is to explore how people from different cultures respond to overt online aggression (flaming) and more covert forms of aggression (ostracism) by ingroup and outgroup members. To do this, we conducted a study across three different countries: Taiwan (representing face-valuing cultures), Pakistan (representing honour-valuing cultures), and the Netherlands to serve as a comparison country. These countries were selected because each had been previously used in extant literature as a representation of one of Leung and Cohen's (2011) three cultural logics: face (Chien et al.,

2018), honour (Anjum et al., 2019), and dignity (Ijzerman & Cohen, 2011). Because of this, we implicitly expected these countries to differ in terms of their people's self-construal (Pakistan and Taiwan having more interdependent people, and the Netherlands having more independent people; see Lee et al., 2014) and their concern for reputation (Pakistan having the most concern, followed by Taiwan, and finally the Netherlands; see Anjum et al. (2019) for a full discussion of how these cultural logics differ in these ways).

In our study, we examined three different types of responses to being either ostracized or flamed in the Cyberball game described in Williams and colleagues' (2000) study: emotional responses, behavioural intentions, and actual behavioural responses. To our knowledge, this is the first study that examines the results of different types of online aggression across all three of our indicators, allowing us to capture nuances in the victim experience that were previously invisible. We were particularly interested in the differences between flaming and ostracism – overt and covert online aggression – in regards to how participants from the three cultural contexts felt (e.g., anger vs. embarrassment), as well as their behavioural intentions; which types of aggression elicit the desire to retaliate in which contexts, for example? We also recorded and coded every message sent by each participant as a record of their behavioural responses to either flaming or ostracism, depending on their assigned condition. In this way, we were able to see not only the practical results of our manipulation, but also how intention differs from action in trolling interactions. Thus, we anticipate the following to occur in the present study:

H1a. When flamed, participants from cultures that value honour will report more anger and be more aggressive than when ostracized (in terms of their intentions and their behaviours), particularly when the perpetrator is an outgroup member as opposed to an ingroup member.

H1b. When ostracized, participants from cultures that value honour will be more likely to feel embarrassed (emotions) than when flamed, and will want to try to repair the relationship (intentions & behaviour), particularly when the perpetrator is an ingroup member as opposed to an outgroup member.

H2a.When flamed, participants from cultures that value face will tend to feel embarrassed and to withdraw compared to when ostracized (in terms of their behavioral intentions and behaviors), particularly when faced with ingroup perpetrators as opposed to outgroup perpetrators.

H2b. When ostracized, participants from cultures primarily valuing face are also likely to feel embarrassed, particularly when ostracized by ingroup members, but they will also be more motivated to try to repair the relationship with them than when they are flamed.

Method

Participants and Design

We conducted the experiment among a sample of 451 participants across the three countries: Taiwan, Pakistan, and the Netherlands. Of these 451 original participants, there were errors saving the Cyberball data of 21, leaving us with 430 participants with completed data. Then, upon further inspection, we noticed that 7 additional participants were below the age of 18, meaning they also had to be removed, leaving us with our final total of 423 participants. The Taiwanese sample consisted of 139 participants (108 women, 31 men) between the ages of 18 and 30 (M = 21.56, SD = 2.36), the majority of whom were highly educated (90), and the rest having obtained a medium level of education (49) according to UNESCO (2011). The Pakistani sample consisted of 149 participants (46 women, 103 men) between the ages of 18 and 41 (M =

22.73, *SD* = 2.95), their education levels evenly split between high (74) and medium (75). The Dutch sample consisted of 135 participants (94 women, 41 men) between the ages of 18 and 27 (*M* = 21.19, *SD* = 2.29), with the majority of these having obtained a low (77) or medium (49) level of education, and only a few (9) having obtained a high level of education (see UNESCO, 2011 for full descriptions of the education levels here).

The study itself took a 3 (nationality: Taiwan, Pakistan, or Dutch) x 3 (types of trolling: flaming, ostracism, or control) x 2 (perpetrator group membership: ingroup or outgroup) experimental design. Participants were randomly assigned to the trolling and perpetrator group membership conditions using Qualtrics' built-in random participant assignment function. In order to be certain that expected country-level patterns did differ in our samples in the ways extant literature described (see Smith et al., 2017), all participants were assessed for individual self-construal and concern for reputation. University-aged students were selected because this is an age group that is likely to be exposed to trolling regularly (Cook et al., 2018). For our ingroup and outgroup manipulation, we chose to use minority groups within each country as our outgroup (Afghani in Pakistan, Filipino in Taiwan, and Moroccan in the Netherlands). Partially, this was because we wanted to be sure to have an effect, and using minority groups when assessing behavioural intentions and aggression had been successful in existing literature (e.g., Schaafsma & Williams, 2012). The study was approved by two separate institutional review boards: one at a mid-size university in Tilburg (whose assessment was also accepted for the Taiwanese portion of the study), and one at a large university in Pakistan.

Procedure

In Taiwan (Taipei) and Pakistan (Lahore), participants were recruited via online advertisements in university fora and university-specific Facebook groups, as well as via

snowball sampling. In the Netherlands (Tilburg), the majority of participants were recruited via a subject pool. Only when the subject pool was depleted were participants recruited using on-campus advertising. Except for the Dutch students who participated for course credit, participants were compensated with a small monetary token appropriate to each country.

Upon arrival in the lab, participants were told that the purpose of the study was to examine mental visualization across cultures (see Williams et al., 2000), and that the full session would consist of a pre-experiment questionnaire, a simple online game, and a post-experiment questionnaire. They were informed that elsewhere in the university, the same procedure was happening with two other participants with whom they would play an online game during the experiment. After giving their consent to participate, participants were asked to complete a questionnaire prior to starting the game with the other two participants. At this point, the research assistant left, and waited outside the room for the participant's knock, signaling that they were ready to begin the game. When the participant knocked, the researcher would re-enter the room, faking having received a text confirming that the other two participants (who are actually pre-programmed computer players) were ready to enter the game. They would then briefly review the mechanics of the game (how to type messages to other players and how to toss the ball) and leave the room again before the game began.

The game itself – Cyberball, a virtual ball toss game – was embedded into Qualtrics. This is a simulation of a game of catch (see for a more elaborate description Williams et al., 2000). Participants each have a simple avatar who take turns tossing a ball back and forth between three or more players, at least some of whom are pre-programmed to behave in a certain way. In the present study, the three-player version was used, and the participant was the only human player. Each game consisted of 30 throws, and took approximately 3 minutes to play. Upon entering the

game, participants received the same instructions (in their local language – Mandarin for Taiwan, English for Pakistan, and Dutch for the Netherlands, see Appendix B) to imagine that they were playing a real game of catch in a real park, engaging their senses as much as possible to create a detailed mental picture. They were asked to introduce themselves to the other players, and were informed that they could chat with the other players in game and were explained how to do so.

Upon entering the game, the two computer players would introduce themselves, giving their nationality (decrying their ingroup or outgroup status, depending on whether or not they shared the nationality of the participant) and a fake interest, either music or football. In Pakistan, for example, an ingroup perpetrator would introduce themselves as follows: "Hi! My name is Ahmed. I grew up here in Lahore. I'm a big fan of football!" Across all conditions, this would be the format, with the following names and outgroups substituted per country in the outgroup conditions: an Afghani named GulShar in Pakistan, a Pilipino named Danilo in Taiwan, and a Moroccan named Mohammed in the Netherlands. After this, the game would proceed depending on the participant's assigned condition. In the control conditions, there was no further chat from the computer players, and these were programmed to pass the ball randomly between each other and the human participant. In the flaming conditions, participants would be repeatedly insulted by Player 1, who would also periodically insult the other computer player. These insults, focusing on the player's childishness (e.g., "You play like a child") or ineptitude, were pre-tested in each of the participating countries to ensure that they were insulting without being ethically dangerous, and equally offensive in each of the countries (see Appendix B). The ball was passed using the same randomized pattern as in the control conditions. In the ostracism conditions, there were no further messages sent by the computer players, but the ball never left the computer

players' avatars; the human participant never had the opportunity to receive or pass the ball, although they were still able to use the chat function.

After the game, participants were redirected to a final questionnaire, which included the main dependent variables and manipulation checks. Once they had completed the questionnaire, the participant knocked on the door a final time and the research assistant re-entered the room. At this point, the assistant would perform a suspicion check by asking the participant what they thought the study was about and would then debrief and give the participant their participation fee (either a course credit or a monetary token, depending on the country of participation). This debrief consisted of the research assistant explaining the purpose and design of the study, including details regarding how the 'other participants' were in fact pre-programmed computerized confederates. They were also careful to explain the random assignment procedure to ensure that no students felt that they were particularly selected to be either ostracized or flamed. After confirming that participants were unharmed and in an acceptable emotional condition, research assistants offered participants a pamphlet explaining trolling and cyberbullying, as well as providing local mental health resources available.

Materials

All materials were administered in a local language: Mandarin for Taiwanese participants, English for Pakistani participants (the university's formal language was English) and Dutch for Dutch participants. In Pakistan, there was no need for translation, as the original language of all measures and scripts was English. For the Dutch and Mandarin editions, all materials were initially submitted for professional translation. After receiving these translations, teams of two to three bilinguals (English-Dutch or English-Mandarin) went over each item and made any adjustments to the language to make sure it corresponded to the original English.

Finally, these versions were back-translated by other bilinguals, and final adjustments were made by the same team of bilinguals that performed the first check. It was this final triple-checked version that was administered. The original English version is presented in Appendix C.

The pre-experiment questionnaire consisted of several demographic questions, as well as a concern for reputation scale and a self-construal scale, to check whether participants across the three samples really differed on these dimensions.

To measure self-construal, we administered one of the subscales (self-interest versus commitment to others) of the initial version of Vignoles and colleagues' (2016) measure of self-construal, which has been validated in multiple languages across 16 countries, including Mandarin. The higher one's score on this scale, the more interdependent (collectivistic) one's self construal. When first examined, the alphas were very low for the three samples (Taiwan, $\alpha = .50$; Pakistan, $\alpha = .48$; The Netherlands, $\alpha = .52$). Upon further examination, it became clear that the final two items ("I should be judged on my own merit" and "I am comfortable being singled out for praise and rewards"), were actually negatively correlated with the rest of the items in the scale. We suspect that these two items – the only two in the scale that were reverse-coded to measure interdependence/collectivism – were in fact measuring independence/individualism as a separate construct instead of merely the inverse of interdependence. We thus removed these items from our analyses. After this procedure, the alpha's for the difference samples (Taiwan, $\alpha = .72$, Pakistan, $\alpha = .68$, and The Netherlands, $\alpha = .65$) all presented an acceptable reliability.

To measure participants' concern for reputation, we employed a modified version of de Cremer and Tyler's (2005) Concern for Reputation scale (CfR). This scale has been validated in both English- and Italian-speaking populations to date (see Cavazza et al., 2014). To capture all

types of reputation described earlier – honour, face, and dignity – this initial scale was expanded to nine items, three for each reputational construct. Participants were asked to indicate on a scale of 1 (not at all) to 5 (extremely) to what extent they felt these statements applied to them. The alphas were acceptable across the three samples: Taiwan = .73, Pakistan = .62, and The Netherlands = .73). Because there were so few items per dimension, and because reputation is only a part of the honour-face-dignity framework, we chose to treat this as a single measure of concern for reputation, although we included representations reflecting each cultural logic (see Leung & Cohen, 2011).

In the post-experiment questionnaire, we measured participants' emotional responses to the game and their behavioural intentions toward the two computer players who they still believed to be other human participants. To assess participants' emotional responses after the game, we used a modified version of the Discrete Emotions Questionnaire (DEQ; Harmon-Jones, Bastian, & Harmon-Jones, 2016). This is a popular emotional evaluative tool that has been used in several different cultural contexts (see Megías et al., 2011; Yilmaz & Bekaroğlu, 2020). Because we were only interested in a few specific emotions, and also wanted to preserve the engage-disengage paradigm put forth by Kitayama and colleagues (2006), we kept the DEQ's format and instructions, but only presented ten items divided into five two-item subscales: positive disengaging emotions (proud, confident), negative disengaging emotions (angry, mad), positive engaging emotions (happy, cheerful), negative engaging emotions (embarrassed, humiliated), and the general construct of respect (respected, ashamed).

Because our interest was in negative emotions resulting from trolling behaviours, we focused on the negative disengaging and engaging emotion examples only in our analyses. However, it is worth noting that the item "ashamed" from the respect construct correlated with

both negative engaging emotions in Pakistan (embarrassed, r = .57; humiliated, r = .53), and the Netherlands (embarrassed, r = .32; humiliated, r = .59), while correlating with the "humiliated" item in Taiwan (r = .38). The "respected" item from the respect construct correlated with both positive approach emotions near equally in Taiwan (proud, r = .53; confident, r = .53), Pakistan (proud, r = .60; confident, r = .53), and the Netherlands (proud, r = .53; confident, r = .57). The items "angry" and "mad", our disengaging emotions, correlated in all three samples (Taiwan, r = .89; Pakistan, r = .75; The Netherlands, r = .71), while the items "embarrassed" and "humiliated", our engaging emotions, only correlated in the Pakistan (r = .63) and Dutch sample (r = .42). In Taiwan, the correlation was negligible (r = .11). We therefore ran our initial analyses with the two items combined, and then another analysis with them separate to determine which item was driving the effects in our earlier tests.

To assess participants' behavioural intentions, we used an adapted version of a scale used by Schaafsma and Williams (2012) when they aimed to assess aggressive behavioural intentions across cultural groups in the Netherlands – more specifically, contrasting honour- and dignity-valuing cultures. We reduced the scale to six items – two aggressive intentions (ex., "hurt the other players"), two reparative intentions ("have a chat with the other players"), and two withdrawal intentions (ex., "stay away from the other players"). During the translation process, we also shortened the scale from seven to five points, and worded the directions to "indicate … how much you want to" and had 1 = not at all, and 5 = extremely. The items representing withdrawal – "stay away from the other players" and "avoid the other players in real life" – were correlated in all three samples (Taiwan, r = .66; Pakistan, r = .61; The Netherlands, r = .29), although the correlation in the Dutch sample was weaker than in the other samples. The items representing aggression – "hurt the other players" and "swear at the other players" – also

correlated in all three samples (Taiwan, r = .55; Pakistan, r = .73; The Netherlands, r = .66). The items representing relationship building or repairing – "have a chat with the other players" and "meet the other players" – correlated weakly, but still significantly, in Taiwan (r = .24), and strongly in Pakistan (r = .59) and the Netherlands (r = 61).

To verify whether the participants were negatively affected by our trolling manipulation, we concluded the post-experiment questionnaire with four manipulation check questions. These consist of an "I felt …" statement (liked, rejected, humiliated, ridiculed), followed by a scale from 1 (not at all) to 5 (extremely). Participants were instructed to "select the number that best represents the feelings you experienced during the game" for each statement. The items were acceptably reliable in all three samples (Taiwan, α = .78; Pakistan, α = .79; The Netherlands, α = .84).

Participants' messages were also coded as a behavioural measure. From 10 basic codes that were given to each message sent by our participants, we reduced it to three macro-codes: retaliation, reparation, and miscellaneous. The number of messages coded as either retaliation or reparation (attempts to repair the relationship) were used as our behavioural measure in our analyses. The full coding procedure is presented in Appendix A

Results

Analytical Strategy

To examine the effects of the three independent variables on the various dependent variables (emotional reactions, behavioural intentions, and behaviour during the game), we first ran a series of three-way MANCOVAs whereby nationality (Taiwanese, Pakistani, Dutch), trolling type (flaming, ostracism, control), and perpetrator group membership (ingroup, outgroup) were included as the between-subjects factors and either anger or embarrassment

(emotions), aggression, reparation or withdrawal (intentions), or reparation and retaliation (behaviour) as within-subjects factors. Means and standard deviations for each of these according to the between-subject factors listed are presented in Table 1. Gender and age were included as covariates in these analyses, but neither ever produced a significant effect – either on their own (Age: $\eta^2 < .001$, $p > .07$, observed power < .09; Gender: $\eta^2 < .001$, $p > .76$, observed power < .06) or in interactions (Age: $\eta^2 < .007$, $p > .10$, observed power < .38; Gender: $\eta^2 < .001$, $p > .11$, observed power < .36) – and so they were removed from final analyses. A correlation matrix of our between-subjects factors is presented in Table 2. Because neither the behavioural measures (retaliation and reparation), nor all of the behavioural intention measures (only aggression and withdrawal, not the intention for reparation) correlated significantly, we chose to run two MANOVAs - one with emotions (anger/embarrassment) as a within-subjects variable, and one with negative intentions (aggression/withdrawal) – and three ANOVAs with the intention to repair the relationship, behavioural retaliation, and behavioural reparation as their respective dependent variables.

Prior to conducting these analyses, we ran a two-way ANOVA with nationality, trolling type, and perpetrator group membership as the independent variables, to examine whether our manipulations had a discernible effect on how liked or rejected participants felt. We also conducted one-way ANOVAs with nationality as the independent variable to check whether or not our three cultural settings differed significantly in terms of our two cultural variables: self-construal and reputation. In addition, a post-hoc power analysis in G*Power (Faul et al., 2007; 2009) revealed that the study is sufficiently powered to successfully detect effects as small as $f^2 = 0.02$, or 2% of the total variance explained (a small effect).

Table 1.

Descriptive statistics of within-subjects factors and dependent variables.

				Anger			Embarrassment			Aggression			Withdrawal			Reparation**	
Nationality	**Trolling**	**Group***	**N**	*M*	*SD*	CI	*M*	*SD*	CI	*M*	*SD*	CI	*M*	*SD*	CI	*M*	*SD*
Taiwanese	None	In	25	1.46	1.04	[1.03, 1.89]	2.40	0.63	[2.14, 2.66]	1.22	0.76	[0.90, 1.54]	1.78	0.97	[1.38, 2.18]	2.65	0.83
		Out	23	1.20	0.49	[0.98, 1.41]	2.37	0.80	[2.02, 2.72]	1.07	0.23	[0.97, 1.16]	1.76	0.67	[1.47, 2.05]	2.82	0.91
	Ostracism	In	21	2.64	1.16	[2.11, 3.17]	3.21	1.04	[2.74, 3.69]	1.57	0.87	[1.18, 1.97]	2.45	1.30	[1.86, 3.05]	3.00	0.81
		Out	20	2.55	1.18	[2.00, 3.10]	3.38	1.06	[2.88, 3.87]	1.65	0.69	[1.33, 1.97]	2.90	1.12	[2.38, 3.42]	3.17	1.06
	Flaming	In	25	2.86	0.90	[2.49, 3.23]	3.48	0.90	[3.11, 3.85]	1.74	0.95	[1.35, 2.13]	3.40	0.91	[3.02, 3.78]	2.98	1.04
		Out	25	2.78	1.23	[2.27, 3.29]	3.56	0.65	[3.29, 3.83]	1.92	1.06	[1.48, 2.36]	2.98	0.99	[2.57, 3.39]	2.60	0.68
Pakistani	None	In	26	1.79	0.96	[1.40, 2.18]	1.52	0.66	[1.25, 1.78]	1.64	0.86	[1.29, 1.98]	1.87	1.03	[1.45, 2.28]	2.76	1.30
		Out	25	1.72	0.82	[1.38, 2.06]	1.30	0.46	[1.11, 1.49]	1.50	0.72	[1.20, 1.80]	2.02	0.96	[1.62, 2.42]	3.29	1.24
	Ostracism	In	24	2.77	1.22	[2.26, 3.28]	2.69	1.14	[2.21, 3.17]	1.83	1.03	[1.40, 2.27]	3.04	1.40	[2.45, 3.63]	2.93	1.18
		Out	23	2.52	1.29	[1.36, 2.32]	2.24	1.20	[1.72, 2.76]	2.20	1.28	[1.64, 2.75]	2.72	1.20	[2.20, 3.24]	3.02	1.15
	Flaming	In	25	1.84	1.15	[1.86, 2.75]	1.82	1.03	[1.40, 2.25]	1.62	0.78	[1.30, 1.94]	2.04	0.92	[1.66, 2.42]	2.60	0.88
		Out	26	2.31	1.11	[1.05, 1.57]	1.83	0.85	[1.48, 2.17]	1.89	1.03	[1.47, 2.30]	2.90	1.17	[2.43, 3.38]	3.30	0.89
Dutch	None	In	21	1.31	0.58	[1.14, 1.77]	1.86	0.71	[1.53, 2.18]	1.10	0.34	[0.94, 1.25]	1.50	0.61	[1.22, 1.78]	2.48	0.96
		Out	22	1.46	0.71	[1.77, 2.56]	1.96	1.05	[1.49, 2.42]	1.23	0.53	[0.99, 1.46]	1.77	0.97	[1.34, 2.20]	2.50	1.10
	Ostracism	In	24	2.17	0.94	[1.69, 2.54]	3.35	0.84	[3.00, 3.71]	1.67	0.82	[1.32, 2.01]	2.69	0.96	[2.28, 3.10]	2.68	1.03
		Out	22	2.11	0.96	[1.53, 2.24]	3.21	0.87	[2.82, 3.59]	1.46	0.58	[1.20, 1.71]	2.61	1.20	[2.08, 3.15]	2.35	1.10
	Flaming	In	22	1.87	0.80	[1.58, 2.55]	2.59	0.93	[2.18, 3.01]	1.52	0.63	[1.25, 1.80]	2.93	1.26	[2.38, 3.49]	2.07	1.05
		Out	24	2.06	1.15	[1.98, 2.19]	2.71	1.23	[2.19, 3.23]	1.56	0.71	[1.26, 1.86]	2.65	1.26	[2.11, 3.18]	2.36	1.01

Note. * Troll's group membership (in-group or out-group)

Table 2.

Correlation matrix of all dependent variables.

	Anger	**Embarrassment**	**Aggression**	**Withdrawal**	**Reparation (I)**	**Retaliation**	**Reparation**
Anger	**1.00**						
Embarrassment	**.56**	**1.00**					
Aggression	**.46**	**.22**	**1.00**				
Withdrawal	**.48**	**.43**	**.45**	**1.00**			
Reparation (I)	**.02**	-.07	.07	**-.21**	**1.00**		
Retaliation	**.16**	.08	**.18**	**.19**	.01	**1.00**	
Reparation	**.11**	.07	**.14**	.04	**.25**	**-.25**	**1.00**

Note. Reparation (I) refers to the behavioural intent to repair the relationship, while Reparation refers to the actual behaviour of sending a message to repair/build the relationship. Bolded correlations are significant at the .05 level.

Preliminary analyses

Our initial ANOVA revealed that both trolling type ($F(2,421) = 90.12$, $p > .001$, $\eta^2 = 0.30$) and nationality ($F(2,421) = 13.07$, $p > .001$, $\eta^2 = 0.06$) had significant effects on our manipulation questions (liked and rejected feelings), but that there was also a significant interaction between these two predictors, $F(4,421) = 7.13$, $p > .001$, $\eta^2 = 0.07$. Participants felt more rejected when flamed or ostracized (means ranged from 2.51 to 3.69) than when in a control condition (means ranged from 2.06 to 2.16), showing that our primary manipulation was successful across all samples tested. In Taiwan, participants felt the most rejected in the flaming condition ($M = 3.69$, $SD = 0.73$), while in Pakistan ($M = 3.20$, $SD = 1.07$) and the Netherlands ($M = 3.36$, $SD = 0.83$), participants felt the most rejected in the ostracism condition. The perpetrator's group membership did not significantly predict our manipulation check results, $F(1,421) = 0.01$, $p = .90$, $\eta^2 < 0.001$, observed power = .05.

An ANOVA also confirmed that there are significant differences between the countries in terms of both self-construal ($F(2,421) = 8.97$, $p < .001$, $\eta^2 = 0.04$) and concern for reputation, $F(2,422) = 48.91$, $p < .001$, $\eta^2 = 0.19$. For self-construal, a Tukey's honest significant difference test revealed that Taiwanese ($p = .04$, $d = 0.31$) and Pakistani participants ($p < .001$, $d = 0.50$) were significantly more interdependent than Dutch participants. There was no significant difference in this regard between the Taiwanese and Pakistani participants, $p = .19$, $d = 0.20$. For reputation, we found that Taiwanese participants were the most concerned about their reputation, followed by Pakistani participants, with the Dutch being the least concerned, all ps $< .001$, Taiwanese to Pakistani, $d = 0.60$; Taiwanese to Dutch, $d = 1.20$; Pakistani to Dutch, $d = 0.57$. These findings confirm the idea that people from face and honour settings see themselves as

more interdependent and also tend to be more concerned about their reputation than those from dignity settings.

Emotional Responses to Flaming and Ostracism

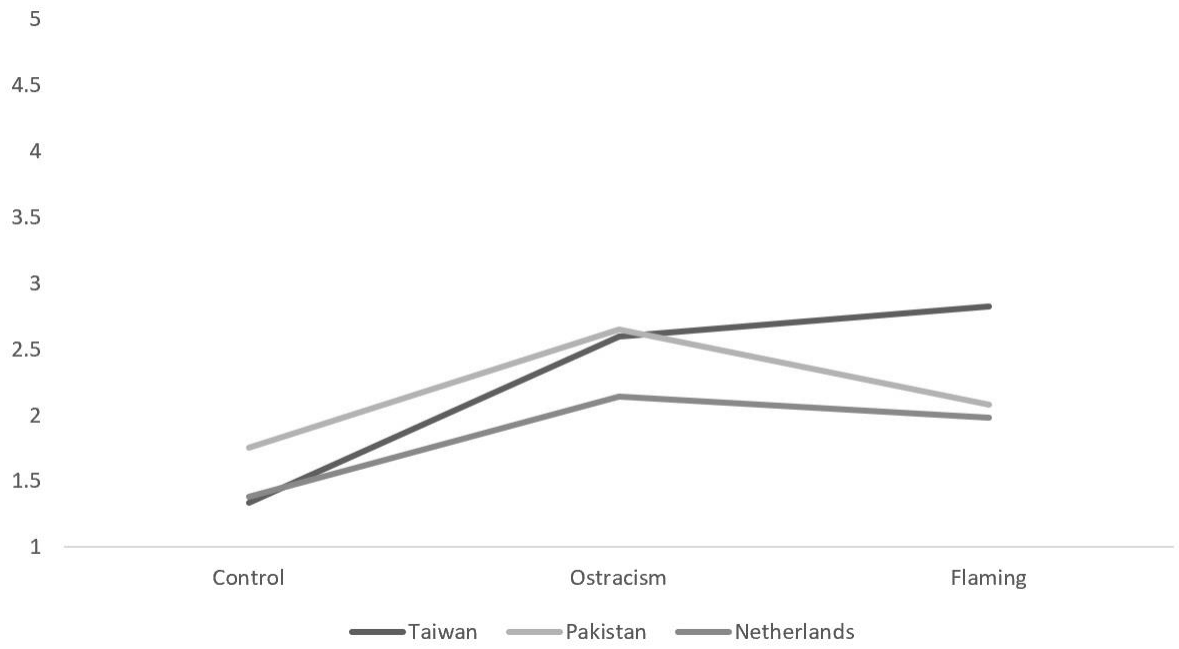
We expected that participants from a culture that values honour (Pakistan) would experience more anger when flamed by outgroup members than by ingroup members (Hypothesis 1a) and feel more embarrassed when ostracized by ingroup members than by outgroup members (Hypothesis 1b). For participants from a face-valuing culture (Taiwan), we anticipated that they would feel more embarrassment when flamed or ostracized by ingroup members than by outgroup members (Hypothesis 2a and 2b, respectively). However, we did not find three-way interactions between nationality, trolling type, and group membership when it comes to negative emotions, $F(4,405) = 0.12$, $p = .98$, $\eta^2 = .001$, observed power = .08.

The analyses did reveal significant main effects of nationality, trolling type, and type of emotion (engaging or disengaging), and a significant interaction between nationality and type of emotions and nationality and trolling type (Fs > 4.85, ps < .001). This last two-way interaction between nationality and trolling type is visualized in Figure 1 (anger) and Figure 2 (embarrassment). Simple effects for this interaction revealed that the difference between the flaming and ostracism conditions in Taiwan ($F(1,420) = 26.38$, $p < .001$, $\eta^2 = .06$), Pakistan ($F(1,420) = 25.97$, $p < .001$, $\eta^2 = .06$), and the Netherlands ($F(1,420) = 26.79$, $p < .001$, $\eta^2 = .06$) were all significant. It would thus appear that Pakistani participants experienced the most anger *and* embarrassment while ostracized, running counter to our expectation that flaming would produce primarily anger (H1a) and ostracism primarily embarrassment (H1b), while Taiwanese participants experienced the most anger *and* embarrassment while flamed, contrary to our

supposition that they would experience primarily embarrassment across both trolling types (H2a and H2b).

Figure 1.

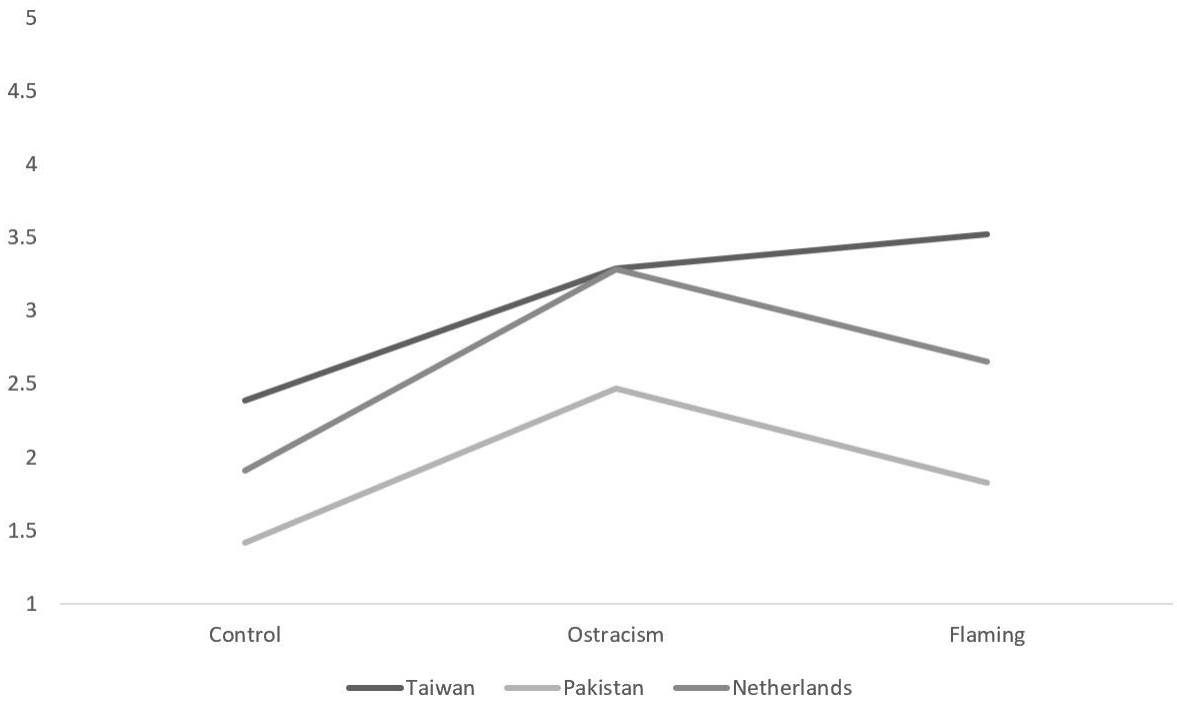
The two-way interaction between nationality and trolling type for anger.



This two-way interaction is given additional nuance, however, from a significant interaction between nationality, trolling type, and type of emotion, $F(4, 405) = 3.39$, $p = .01$, $\eta^2 = .03$. The simple effects tests that we conducted to examine this revealed a significant two-way interaction between negative emotions and trolling for the Dutch sample only, $F(2, 416) = 4.56$, $p = .01$. A further inspection of this interaction revealed that they experienced significantly more embarrassment than anger in the control ($F(1, 420) = 10.92$, $p = .001$), ostracism ($F(1, 420) = 55.56$, $p < .001$), and flaming conditions ($F(1, 420) = 19.37$, $p < .001$).

Figure 2.

The two-way interaction between nationality and trolling type for embarrassment.



Given that the two engaging negative emotion items (embarrassment and humiliation) were not correlated in the Taiwan sample, we conducted two additional ANOVAs with the same between-subjects variables as before, and the two items, "embarrassed" and "humiliated", as dependent variables. These analyses revealed that in the Taiwan sample, the mean levels of embarrassment were similar in the trolling and non-trolling conditions ($F(2,414) = 1.94$, $p = .15$), but there were significant differences in humiliation between the flaming and ostracism conditions, $F(1,414) = 83.46$, $p > .001$. Flaming ($M = 3.64$, $SD = 1.05$) resulted in more humiliation than ostracism ($M = 2.61$, $SD = 1.38$), $F(1,414) = 23.06$, $p < .001$, and these two trolling conditions resulted in more humiliation than the control condition ($F(1,414) = 83.46$, $p > .001$). Overall, these results would suggest that the effect we found earlier – the lack of

difference between flaming and ostracism– is because of the embarrassment item, as we do find a difference between the two in terms of humiliation.

Behavioural Intentions toward Perpetrators of Flaming and Ostracism

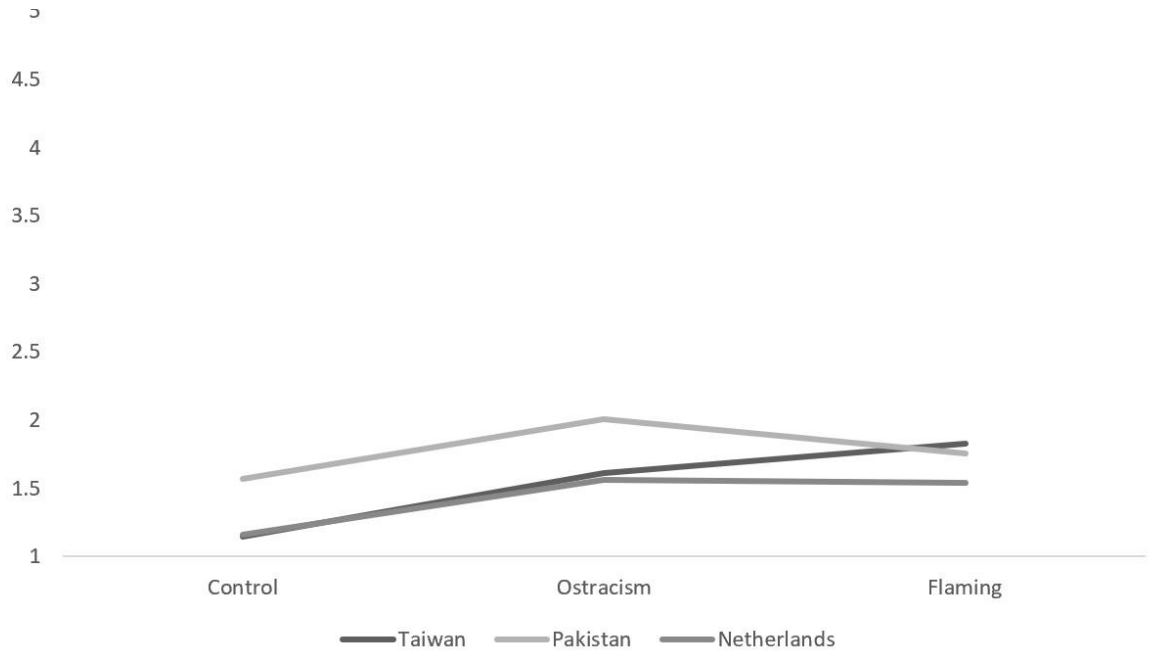
In terms of behavioural intentions, we expected that our honour-valuing culture participants would express more aggressive intentions, particularly when flamed by outgroup members (Hypothesis 1a), but that they want to try to repair the relationship when ostracized, particularly by ingroup member (Hypothesis 1b). For our face-valuing culture participants, we anticipated that flaming (by ingroup members in particular) would lead them to withdraw, but that ostracism by ingroup members would lead to participants trying to restore their relationship with the perpetrator (Hypothesis 2b). It is important here to note that aggressive intention scores were low across all cultures; none of our sample could be considered truly 'aggressive' in their responses.

For the negative behavioural intentions, our analysis revealed significant main effects of trolling type and negative intentions, as well as significant interactions between, nationality and negative intentions, trolling type and negative intentions, and another between nationality and trolling type (Fs > 4.77, ps < .01). This last interaction is visualized in Figure 3 (aggression) and Figure 4 (withdrawal). Simple effects analyses revealed that the differences in negative behavioural intentions – both anger and withdrawal – between the flaming and ostracism conditions were significant in Taiwan ($F(1,420) = 14.74$, $p < .001$, $\eta^2 = .03$), Pakistan ($F(1,420) = 15.70$, $p < .001$, $\eta^2 = .04$), and the Netherlands, $F(1,420) = 15.07$, $p < .001$, $\eta^2 = .03$. Our Taiwanese participants thus wanted to aggress *and* withdraw the most when flamed, contradicting the idea that withdrawal would be expressed equally between trolling conditions (H2a and H2b). Nevertheless, it does partially support our prediction in H1a that Pakistani

participants would feel heightened anger when flamed, although perpetrator group membership did not play a role in this experience.

Figure 3.

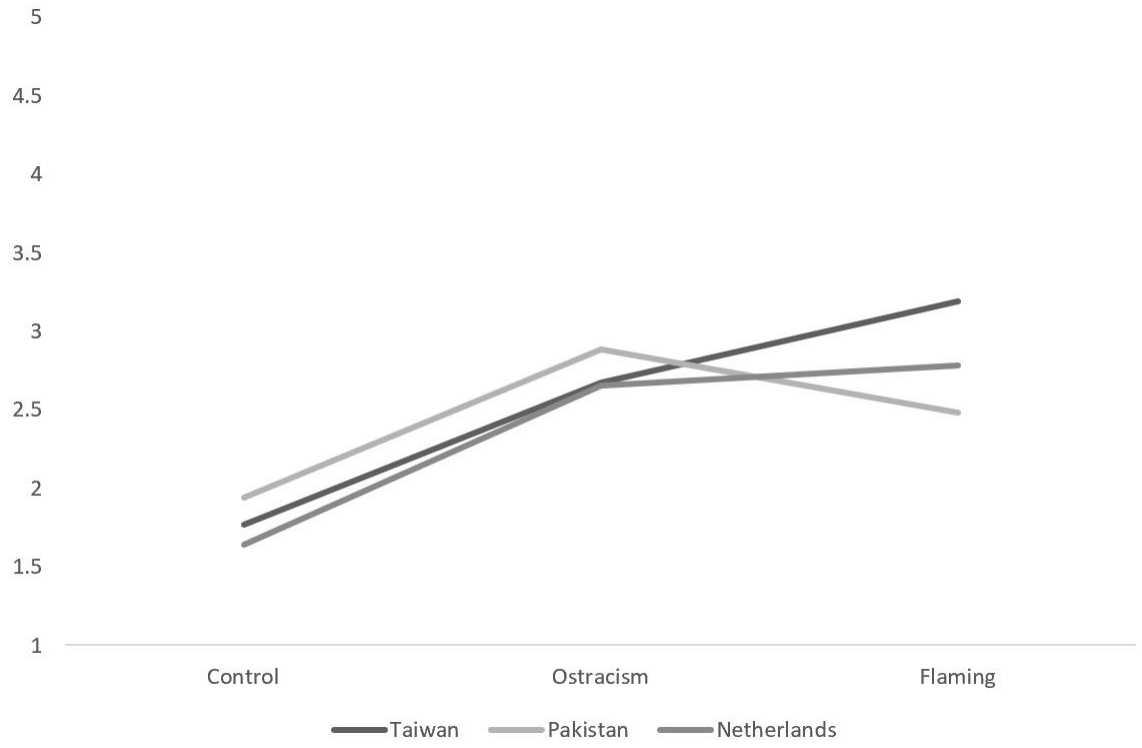
The two-way interaction between nationality and trolling type for aggression.



However, our Pakistani participants were also involved in a within-between effects four-way interaction between nationality, trolling type, troll's group membership, and negative intentions, $F(4,405) = 3.21$, $p = .01$, $\eta^2 = .04$. This interaction was only significant in Pakistan ($F(2,416) = 4.65$, $p = .01$; all other ps > .08). Simple effects tests revealed that in the Pakistani sample, when participants were ostracized ($F(1,419) = 4.63$, $p = .03$) by an in-group perpetrator, they intended to withdraw ($M = 2.72$, $SD = 1.20$) more than aggress ($M = 2.20$, $SD = 1.28$), but when faced with an out-group perpetrator, they intended to aggress ($M = 3.04$, $SD = 1.40$) more than withdraw ($M = 1.83$, $SD = 1.03$), which confirms what we predicted in H1b. No such difference was found for the Taiwan and Dutch participants.

Figure 4.

The two-way interaction between nationality and trolling type for withdrawal.



For reparative intentions – we not only found no hypothesized three-way interaction ($F(4,405) = 1.34$, $p = .26$), but we found no significant main effects (all Fs < 2.02, all ps > .16) or lower-order interactions (all Fs < 2.33, all ps > .10) at all. This would suggest that participants across all countries experienced equal desire (all means ranged from 2.06 to 3.17) to meet and befriend the computer players, irrespective of whether these players trolled them or not. That said, these means are quite low, suggesting that few participants had any desire to actually go and meet the other players or befriend them in person. This also goes against the idea as expressed in hypotheses 1b and 2b that members of face-valuing and honour-valuing cultures would want to repair the relationship with ingroup ostracizers, as neither our Taiwanese nor our

Pakistani sample expressed any major desire to do so, irrespective of trolling condition or perpetrator group membership.

Behavioural Responses to Flaming and Ostracism

Because the actual behavioural data – the messages participants sent during the game – consisted of count data, they could not be analysed using the same techniques as the emotional responses and behavioural intentions. Since our interest was in relative usage (e.g., do Pakistani participants more often retaliate or try to repair the relationship?), we first calculated the proportion of total messages used to either retaliate or repair (in the case of no messages sent, a 0 was entered manually to signify that none of the messages sent pertained to retaliation or reparation) according to our coding scheme (see Table 1 and Appendix A). We then calculated the descriptive statistics for these proportions by nationality, trolling type, and troll's group membership, which are presented in Table 3.

For behavioural aggression (retaliation), we did not find the three-way interaction we anticipated, $F(4, 405) = 0.52$, $p = .72$, $\eta^2 = .005$, observed power = .18. We did, however, find a significant main effect of type of trolling ($F(2,405) = 39.18$, $p < .001$, $\eta^2 = .16$), as well as a significant interaction between nationality and trolling type, $F(4,405) = 8.31$, $p < .001$, $\eta^2 = .08$. No other interactions were significant (all $Fs < 0.52$, $ps > .72$), nor was there an effect of perpetrator group membership, $F(1,405)$ $p = .33$. The significant interaction between nationality and trolling type is visualized in Figure 5. Follow-up analyses for revealed that in Pakistan, participants responded with aggression significantly more when flamed or ostracized than when in the control condition ($F(1,414) = 6.35$, $p = .01$), but that there was no significant difference between the flaming and ostracism condition, $p = .92$. APA This partially supports our expectation that participants from honour-valuing cultures (Pakistan) would be more likely to

Table 3.

Descriptive statistics of participants' behavioural responses to trolling.

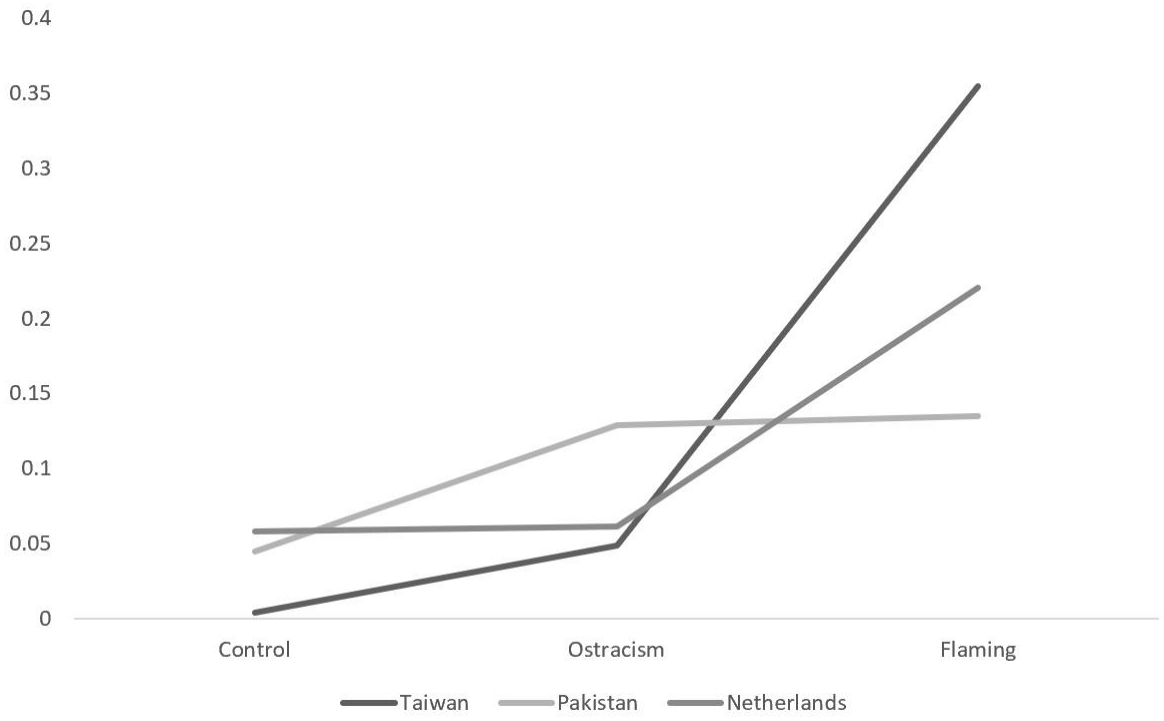
				Retaliation**		**Reparation****	
Nationality	**Trolling**	**Group***	**N**	*M*	*SD*	*M*	*SD*
Taiwanese	None	In	25	0.01	0.04	0.20	0.26
		Out	23	0.00	0.00	0.22	0.29
	Ostracism	In	21	0.08	0.17	0.52	0.28
		Out	20	0.02	0.06	0.53	0.24
	Flaming	In	25	0.33	0.25	0.23	0.25
		Out	25	0.38	0.28	0.25	0.24
Pakistani	None	In	26	0.06	0.18	0.40	0.36
		Out	25	0.03	0.10	0.34	0.41
	Ostracism	In	24	0.14	0.23	0.70	0.26
		Out	23	0.12	0.22	0.55	0.36
	Flaming	In	25	0.13	0.27	0.38	0.38
		Out	26	0.14	0.28	0.34	0.39
Dutch	None	In	21	0.09	0.21	0.18	0.23
		Out	22	0.03	0.14	0.33	0.35
	Ostracism	In	24	0.06	0.16	0.39	0.31
		Out	22	0.06	0.16	0.35	0.30
	Flaming	In	22	0.26	0.28	0.25	0.26
		Out	24	0.19	0.25	0.26	0.30

Note. * Troll's group membership (in-group or out-group). ** Proportion of total messages.

react aggressively following flaming from outgroup members. We found that Taiwanese participants, however, were more likely to aggress when trolled (flamed or ostracized) compared to the control condition ($F(1,414) = 29.63$, $p < .001$) and also more in the flaming than in the ostracism condition, $F(1,414) = 53.22$, $p < .001$. We found the same pattern among our Dutch participants, who also reacted with aggression more when trolled than when in a control condition ($F(1,414) = 4.91$, $p = .03$), and more when flamed than when ostracized, $F(1,414) = 14.33$, $p < .001$. This again goes against the idea that face-valuing culture members avoid aggression due to rejection avoidance tendencies (e.g., Hashimoto & Yamagishi, 2013; 2016), as our Taiwanese participants retaliated in both our ostracism and flaming conditions.

Figure 5.

The two-way interaction between nationality and trolling type for retaliation.



In terms of reparation, we expected that honour-valuing and face-valuing participants would try to repair the relationship, particularly when ostracized by ingroup members. We did not find, however, the hypothesized three-way interaction ($F(4,405) = 0.38$, $p = .82$, $\eta^2 = .003$, observed power = .14), nor did we find any lower-order interactions (all Fs < 1.69, all ps > .19). Instead, we found two significant main effects: one of nationality ($F(2,405) = 10.24$, $p < .001$, $\eta^2 = .05$), and the other of trolling type, $F(2,405) = 23.90$, $p < .001$, $\eta 2 = .11$. A series of simple contrasts revealed that while Pakistani participants sent proportionately more reparation messages than both the Taiwanese ($F(1,414) = 13.00$, $p < .001$) and Dutch ($F(1,414) = 16.97$, $p < .001$) participants, there was no significant difference between the amount of reparation expressed by the Taiwanese and the Dutch, $p = .59$. Another series of simple contrasts showed

that participants in the ostracism condition expressed proportionately more reparation than participants in the flaming ($F(1,414) = 36.89$, $p < .001$) and control ($F(1,414) = 35.59$, $p < .001$) conditions, and that there was no significant difference in terms of reparation between these latter two, $p = .87$. Thus, we cannot confirm that honour- and face-valuing culture members tend to engage in repairing the relationship when flamed (honour) or ostracized (honour and face) by ingroup members, as we found no interactions between nationality/culture, trolling type, or perpetrator group membership.

Discussion

The aim of this study was to examine how people respond to two different types of trolling – flaming and ostracism – and whether this varies as a function of their cultural background and the perpetrator's group membership – ingroup or outgroup. Based upon previous theorizing and empirical work on honour and its connections with aggression, we expected that participants from honour-valuing cultural contexts would react with anger and aggression to flaming, particularly in the case of outgroup perpetrators, but would feel embarrassed and try to repair the relationship when ostracized, particularly when ostracized by ingroup members. We also expected that our participants from a face-valuing cultural context would generally feel embarrassed and try to withdraw and avoid conflict when faced with flaming, but would try to repair the relationship with ingroup ostracizers.

Our results provided mixed support for the idea that honour-valuing culture members should respond with anger and aggression when flamed by outgroup members (H1a). For example, although we did find that flaming resulted in anger among our Pakistani participants, it actually produced less anger than ostracism and also did not vary as a function of perpetrator group membership. After being flamed, Pakistani participants expressed the desire to withdraw

from instead of aggress the perpetrator, particularly when the perpetrator was an ingroup member. During the flaming, Pakistani participants did react with aggression, but no more so than if they were being ostracized, and irrespective of the perpetrator's group membership. In terms of their reactions to ostracism, we found that they did generally withdraw emotionally and experience embarrassment when faced with ostracism, also irrespective of the perpetrators group membership, although no more than if they were faced with flaming (H1b). After being ostracized, they expressed the desire to aggress outgroup perpetrators and withdraw from in-group perpetrators though, which was consistent with our predictions. Pakistani participants were also the most likely to engage in reparation, as were participants in ostracism conditions, but there was no effect of perpetrator group membership. This final result (high rates of reparation) is likely due to the importance of reparation after conflict in Pakistani culture (Anjum et al., 2017). Although research would suggest that an action is preferred over words (Anjum et al., 2018), words are all that the present experimental paradigm allowed, and so it was the medium of reparation our Pakistani participants used.

Taken together, these findings go against the idea presented in literature that the presence of an obvious insult is what triggers aggression in honour-valuing culture members (e.g., Harinck et al., 2013), as reactions from our Pakistani participants to ostracism and flaming only differed slightly. Whether flamed or ostracized, Pakistani participants expressed anger and were equally likely to retaliate, despite the fact that flaming consists of direct insults (O'Sullivan & Flanagin, 2003), while ostracism's insult is implicit and left open to interpretation (Williams, 2009). This also appears to go against Pfundmair and colleagues' (2015) findings that aggressive intentions follow ostracism in more collectivistic cultures. Our results would suggest that flaming and

ostracism are equally threatening when it comes to a loss of honour; both are interpreted as being insulting and requiring of defence.

Our null findings when it came to perpetrator group membership (ingroup or outgroup) in the flaming conditions may have to do with, at least in Pakistan's case, our Pakistani participants' relationships with other Pakistani university students (the ingroup) and Afghani people (the outgroup). Research has shown that when it comes to intergroup communication and conflict among Pakistani people, a key mechanism behind the hostility, emotionally and intentionally, is group relative deprivation (Obaidi et al., 2019). In essence, if the victim feels underprivileged financially or socially compared to the other group, this will trigger a more extreme reaction emotionally and in terms of behavioural intentions. Although Pakistani participants did sometimes react with anger and hostile intentions, means were very low across all reactions, and withdrawal was a much more popular option overall; this could suggest that our Pakistani participants felt relatively equal to both other students and to Afghani migrants. It could also be that there are fundamental differences between cultural values' application with strong and weak ties, as discussed earlier in our implicit research question. Further research could confirm or deny this possibility by performing a replication that included socio-economic data, something that was not collected in the present study, or by directly manipulating the strength of a troll's tie to the victim/participant.

Another possible explanation for these results could be due to the way that honour is conceived in different cultural contexts. Honour is typically shared by the wider social group (Leung & Cohen, 2011), and unprovoked insults and wilfully ignoring people are both likely in violation of local honour codes (e.g., Anjum et al., 2019; Cohen et al., 1996). Participants may have felt like they needed to defend the honour of their culture and homeland – thus

consequently their own honour – by correcting this perceived misrepresentation of what is allowable in their country (e.g., Anjum et al., 2019; Rodriguez-Mosquera et al., 2008). Though outgroup perpetrators do not share a nationality with the participant in the present experiment, they do reside in the same city, giving participants a reason to want them to act in accordance with local norms and properly represent their ingroup. From this perspective, it is not the presence or absence of an overt insult that creates retaliation, but rather the transgression of norms, rendering unwarranted flaming and ostracism equally reprehensible. This would require further in-depth research in other honour-valuing cultural contexts to confirm or deny, but the present study does make the validity of overt insult as a mechanism for reactive aggression uncertain.

Our finding that ostracism and flaming are both considered equally offensive in the Pakistani context is, however, in line with research on negative self-conscious emotions (e.g., Allpress et al., 2014; Giner-Sorolla, 2019; Prati & Giner-Sorolla, 2018), and would suggest that both flaming and ostracism present a threat to a person's social identity (Chen et al., 2020; Dasborough et al., 2020). This also suggests that physical and verbal aggression is used to rebuff minor social infractions in honour-valuing cultural contexts primarily when face-to-face (Harinck et al., 2013; Severance et al., 2013). If this is indeed the case, our results may mean that the online context somehow levels the playing field and makes overt and covert aggression equally hurtful. This would, however, require further research to confirm or deny, and if it is indeed the case, the exact mechanism behind this effect remains unclear.

In terms of our expectation that participants from a face-valuing culture would be embarrassed and withdraw when faced with ingroup flaming (H2a), we again found only partial support. Emotionally, our Taiwanese did indeed experience embarrassment when flamed, which

is in line without our expectations and also self-conscious emotion research (Allpress et al., 2014; Dasborough et al., 2020). However, there was no evidence to suggest that they intend to withdraw after being flamed any more than they do after being ostracized, and instead of trying to repair the relationship with the perpetrator, Taiwanese participants were likely to retaliate against their aggressor, irrespective of that person's group membership. We also expected our face-valuing culture participants to feel embarrassed and want to try to repair the relationship with ingroup ostracizers (H2b), but this was also not fully supported by our results. Although our Taiwanese participants did feel embarrassed when ostracized, they seemed to have mixed or uncertain intentions toward the perpetrator after the fact, and did not appear to retaliate any more than they engaged in reparation during the game. Once more, we found no effect of perpetrator group membership.

These results are surprising – especially the unexpectedly high retaliation and aggression rates among Taiwanese participants – as they appear to contradict the vast majority of the literature on face-valuing cultures, although they are in line with studies that focus on honour in traditionally face-valuing cultural contexts (e.g., Anjum et al., 2019). Cross-cultural studies often paint face-valuing cultures as being bent on rejection avoidance (e.g., Hashimoto & Yamagishi, 2013; 2016). Their bottom line is fitting in and avoiding stirring up conflict, as aggressive conduct is considered shameful and is likely to result in a loss of face (Leung & Cohen, 2011) for the person and their close others (see Markus & Kitayama, 1991). However, the context in which this study takes place – the internet – could provide a theoretical explanation for our findings in Taiwan. Just as our participants have a set of norms to which they generally adhere in their daily life – what we call their culture – the internet itself also has its own social norms (see Phillips, 2016). Our participants are taking their own cultural values into a unique culture in and

of itself when they go online. One thing that has been repeatedly demonstrated about the internet's global culture is that in it, trolling in all its forms is exceedingly common (see Cook et al., 2018, Phillips, 2016). By retaliating against flaming, although our participants are contravening the norms of their own culture, they are actually fitting in to the internet's culture.

Another possible explanation of our results could be the anonymity of the internet and the perceived closeness of our Taiwanese participants to their fellow Taiwanese students. Online, no one can be sure of who you are (see Postmes et al., 1998), which means that no one can associate what you do with any of your close others. While in an offline situation, a person from a face-valuing cultural context would be risking a loss of face for themselves and their ingroup (Leung & Cohen, 2011; Hashimoto & Yamagishi, 2013; 2016), anonymity can prevent that loss of face entirely, as no one could connect their aggression to their group. It could be that the online context simply gave our participants the freedom to give knee-jerk reactions instead of having to consider any face-related consequences. It is also possible that even in face-valuing cultures, the direct insults make honour concerns more salient in the present study than face concerns, hence the results being more in line with Anjum and colleagues' (2019) study as opposed to the work of Hashimoto and Yamagishi (2013; 2016).

Finally, we must consider our results in the dignity-valuing context. Though we had no specific hypotheses regarding our Dutch participants, they were intended to act as a comparison group, a representation of dignity-valuing (Leung & Cohen, 2011), independent (Markus & Kitayama, 1991) culture. Again, our results here were not in line with previous literature. Instead of feeling primarily anger (Rodriguez Mosquera et al., 2008), Dutch participants felt mostly embarrassment when trolled, and there were no distinctions between ostracism and flaming. They also retaliated the least of our three samples, irrespective of trolling type, when extant

literature suggests that a more independent self-construal usually leads to retaliation (see Ma & Bellmore, 2016). From a theoretical standpoint, this is difficult to explain. It is possible that the anonymity of the internet removed the urgency from the situation; it is much harder to ignore people insulting a person to their face as opposed to from behind a screen (see Kyom, 2016). Future studies could explore this idea by explicitly measuring the importance of the medium to participants' lives and communication.

Limitations

Despite our intriguing results, this study is not without its limitations. First among these is our sample of university students. Although they fit into the age range of some of the heaviest internet users and trolls (see Cook et al., 2018), using university students for experiments comes with its own risks. Across all countries, we had a minimum of 50% successful guess rate when we asked participants what they thought the study was about, despite our cover story. This is likely to be because all three participating universities had some form of computer science program in which artificial intelligence and chat-bots were featured. Although every effort was made to make it look like real people were playing, the salience of chat-bots among the student populations tested cannot be denied. Although research has found that, even when participants are aware of a perpetrator being a machine, it does not change the negative effects of online hostility (Zadro et al., 2004), the ecological validity would have been boosted significantly if fewer participants guessed the study's true purpose. There may also be a question of power, as the effect sizes in the present study were notably small; future studies should aim to have even more participants to detect even smaller effects than we were able to in the present work. We are also unsure of how salient our ingroup/outgroup manipulation was, and this could have also contributed to the lack of results when it came to that variable. Beyond this, our sample was

relatively small for a cross-cultural study, and would have been more powerful with additional participants, preferably from a variety of universities within the countries in question to compensate for participants' potential familiarity with AI agents like chat-bots. Future studies should actively take media experience and technological familiarity into account, even when the primary interest is in cultural effects. More specifically, gaming experience should be measured, as this could explain some of the effects we found in the present experiment.

One final important limitation of the present study is the potential confound inherent to the study design when it comes to disentangling ostracism and flaming. Because we used pre-programmed confederates for consistency, natural responses to participants' messages or inquiries were not possible. This means that in both the control conditions and flaming conditions, participants did experience a form of ostracism (their messages receiving no response), albeit not as total as the one they experienced in the actual ostracism conditions. In addition, although it was intended to serve as a passive bystander – something that is seen quite regularly in online contexts (see Cook et al., 2018) – in the flaming condition, the 'bystander' confederate also served as a co-victim, as the troll confederate periodically insulted them as well, while in the ostracism condition, they served as co-troll, as they did not address the participant either. Although it is evident that the key element of verbal insult was unique to the flaming conditions, and keeping the ball away from the participant was unique to the ostracism conditions, some forms of ostracism did likely bleed through all conditions. Thus, while we still found significant differences between the types of trolling in terms of the emotional response, intentions, and behavioural response, future studies performing comparisons of this kind should be extremely careful to ensure that they are fully separate. If they intend to use pre-programmed confederates as we did, advances in natural language processing might make this easier, while

also boosting the ecological validity. Further branching scripts with human confederates may also help in this endeavour.

Conclusions & Future Directions

So which is worse: ostracism or flaming? Our results do not offer a firm conclusion, but rather a resounding "it depends". Emotionally, it would seem that flaming is a more intense experience for people from face-valuing cultural contexts, while ostracism is more intense for people from honour- and dignity-valuing cultural contexts. If "worse" is defined as producing more aggression, then flaming would be worse for face-valuing people, while ostracism would be worse for honour-valuing people. Dignity-valuing people produced so little aggression and retaliation when faced with either type of trolling that the two seem about even. Still, despite not giving a concrete answer to the question of which is worse, overt or covert aggression, the present study has advanced our understanding of both types of aggression in the online context in several ways. While Zadro and colleagues (2004) were among the first to compare ostracism and verbal aggression (flaming), our results expand upon their findings by looking at multiple indicators beyond the traditional effects (senses of belonging, control, self-esteem, and a meaningful existence) of ostracism. While earlier studies confirmed that responses to ostracism differ between cultural groups (e.g., Uskul & Over, 2014; Garris, Ohbuchi, Oikawa, & Harris, 2011), the present study revealed that this is not just true of behavioural responses, but also emotional and intentional responses to not just ostracism, but also flaming, which has only limited cross-cultural studies in its extant literature (e.g., De Seta, 2013).

The study also joins the growing scholarship on online aggression, and opens up the question of how much influence the medium has on responses to both overt and covert aggression. In our dignity-valuing sample, for instance, flaming and ostracism were near equal in

their effects: is this because of some mechanism related to being online, or something else? Extant literature frequently posits that dignity-valuing people are also the most provocative and retaliatory in their responses to aggression (e.g., Ma & Bellmore, 2016), but they were the least aggressive of all of our samples. Future studies should explore this further, manipulating not only the subtlety of the aggression, but also the medium, in order to isolate these sorts of effects. This type of study should also be conducted again in several cultural contexts, as there appear to be effects in non-honour-valuing cultures that are as of yet uncovered by existing theory (e.g., Dutch participants experiencing high levels of embarrassment) that could be specific to our samples. There is still much work to be done, and great opportunities for CMC, aggression, and cultural scholars to collaborate and explore this newest arena of hostility and intercultural communication.

References

Allpress, J.A., Brown, R., Giner-Sorolla, R., Deonna, J.A., & Teroni, F. (2014). Two faces of group-based shame: Moral shame and image shame differentially predict positive and negative orientations to ingroup wrongdoing. *Personality and Social Psychology Bulletin, 40*, 1270-1284. doi: 10.1177/0146167214540724

Ang, R.P., Huan, V.S., & Florell, D. (2014). Understanding the relationship between proactive and reactive aggression, and cyberbullying across United States and Singapore adolescent samples. *Journal of Interpersonal Violence*, *29*, 237-254. DOI: 10.1177/0886260513505149

Anjum, G., Castano, E., & Aziz, M. (2017). Reparations to the victims of the US drone strikes: Youth perspective from Pakistan. In S.S. Aneel, U.T. Haroon, & I. Niazi (Eds.), *Securing Peace and Prosperity* (pp. 245-249). Lahore, Pakistan: Sang-e-Meel Publications.

Anjum, G., Kessler, T., & Aziz, M. (2019). Cross-cultural exploration of honor: Perception of honor in Germany, Pakistan, and South Korea. *Psychological Studies*, *64*, 147-160. doi: 10.1007/s12646-019-00484-4

Anjum, G., Kidd, D.C., & Aziz, M. (2018). Hope in times of terrorism: Action-expressions speak louder than passive-sorrows. *Journal of Behavioral Sciences*, *28*(1), 1-17. Retrieved from http://pu.edu.pk/images/journal/doap/PDF-FILES/01_v28_1_18.pdf

Arjoranta, J., & Siitonen, M. (2018). Why do players misuse emotes in Hearthstone? : Negotiating the use of communicative affordances in an online multiplayer game. *Game Studies: The International Journal of Computer Game Research, 18*. Retrieved from http://gamestudies.org/1802/articles/arjoranta_siitonen

Bond, M.H., Wan, K.-C., Leung, K., & Giacalone, R.A. (1985). How are responses to verbal insult related to cultural collectivism and power distance? *Journal of Cross-Cultural Psychology*, *16*, 111-127. Retrieved from https://journals.sagepub.com/doi/pdf/10.1177/0022002185016001009

Bresnahan, M.J., Shearman, S.M., Lee, S.Y., Ohashi, R., & Mosher, D. (2002). Personal and cultural differences in responding to criticism in three countries. *Asian Journal of Social Psychology*, *5*, 93-105. DOI: 10.1111/1467-839X.00097

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

Cavazza, N., Pagliaro, S., & Guidetti, M. (2014). Antecedents of concern for personal reputation: The role of group entitativity and fear of social exclusion. *Basic and Applied Social Psychology*, *36*, 365-376. doi: 10.1080/01973533.2014.925453

Chen, Y., Li, L., Ybarra, O., & Zhao, Y. (2020). Symbolic threat affects negative self-conscious emotions. *Journal of Pacific Rim Psychology*, *14*, 1-8. doi: 10.1017/prp.2020.3

Chien, S.-Y., Lewis, M., Sycara, K., Liu, J.-S., & Kumru, A. (2018). The effect of culture on trust in automation: Reliability and workload. *ACM Transactions on Interactive Intelligent Systems*, *8*(4), Article 29. doi: 10.1145/3230736

Cohen, D. (1998). Culture, social organization, and patterns of violence. *Journal of Personality and Social Psychology*, *75*, 408-419. Retrieved from http://web.b.ebscohost.com.tilburguniversity.idm.oclc.org/ehost/pdfviewer/pdfviewer?vid=4&sid=b695c456-369c-4451-9ad1-c9738ffa4fcd%40pdc-v-sessmgr02

Cohen, D., & Nisbett, R.E. (1997). Field experiments examining the culture of honour: The role of institutions in perpetuating norms about violence. *Personality and Social Psychology*

Bulletin, *23*, 1188-1199. Retrieved from https://journals-sagepub-com.tilburguniversity.idm.oclc.org/doi/pdf/10.1177/01461672972311006

Cohen, D., & Nisbett, R.E. (1994). Self-protection and the culture of honour: Explaining southern violence. *Personality and Social Psychology Bulletin: The self and the collective*, *20*, 551-567. Retrieved from https://journals-sagepub-com.tilburguniversity.idm.oclc.org/doi/pdf/10.1177/0146167294205012

Cook, C., Conijn, R., Antheunis, M.L., & Schaafsma, J. (2019). For whom the gamer trolls: A study of trolling interactions in the online gaming context. *Journal of Computer-Mediated Communication*. Advance online publication. DOI: 10.1093/jcmc/zmz014

Cook, C., Schaafsma, J., & Antheunis, M. L. (2018). Under the bridge: An in-depth examination of online trolling in the gaming context. *New Media and Society*, *20*, 3323-3340. DOI: 10.1177/1461444817748578

De Cremer, D., & Tyler, T.R. (2005). Am I respected or not? : Inclusion and reputation as issues in group membership. *Social Justice Research*, *18*, 121-153. DOI: 10.1007/s11211-005-7366-3

Cross, S.E., Uskul, A.K., Gerçek-Swing, B., Alözkan, C., & Ataca, B. (2013). Confrontation versus withdrawal: Cultural differences in responses to threats to honor. *Group Processes & Intergroup Relations*, *16*, 345-362. DOI: 10.1177/1368430212461962

Dasborough, M.T., Hannah, S.T., & Zhu, W. (2020). The generation and function of moral emotions in teams: An integrative review. *Journal of Applied Psychology*, *105*, 433-452. doi: 10.1037/apl0000443

DeMarco, T.C., & Newheise, A.-K. (2018). Coping with group members who insult the in-group. *Social Psychological and Personality Science, 9,* 234-244. doi: 10.1177/1948550617732392

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. Behavior Research Methods, 41, 1149-1160. Retrieved from https://www.psychologie.hhu.de/fileadmin/redaktion/Fakultaeten/Mathematisch-Naturwissenschaftliche_Fakultaet/Psychologie/AAP/gpower/GPower31-BRM-Paper.pdf

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods, 39, 175-191. Retrieved from https://www.psychologie.hhu.de/fileadmin/redaktion/Fakultaeten/Mathematisch-Naturwissenschaftliche_Fakultaet/Psychologie/AAP/gpower/GPower3-BRM-Paper.pdf

Filipkowski, K.B., & Smyth, J.M. (2012). Plugged in but not connected: Individuals' views of and responses to online and in-person ostracism. *Computers in Human Behavior, 28,* 1241-1253. doi: 10.1016/j.chb.2012.02.007

Freedman, G., Williams, K.D., & Beer, J.S. (2016). Softening the blow of social exclusion: The responsive theory of social exclusion. *Frontiers in Psychology, 7,* 1-17. DOI: 10.3389/fpsyg.2016.01570

Garris, C. P., Ohbuchi, K.-I., Oikawa, H., & Harris, M. J. (2011). Consequences of interpersonal rejection: A cross-cultural experimental study. *Journal of Cross-Cultural Psychology, 42,* 1066-1083. DOI: 10.1177/0022022110381428

Giner-Sorolla, R. (2019). The past thirty years of emotion research: Appraisal and beyond. *Cognition and Emotion, 33,* 48-54. doi: 10.1080/02699931.2018.1523138

Glick, P., Sakalli-Uğurlu, N., Akbaş, G., Orta, I. M., & Ceylan, S. (2016). Why do women endorse honour beliefs? Ambivalent sexism and religiosity as predictors. *Sex Roles, 75,* 543-554. DOI: 10.1007/s11199-015-0550-5

Gonzales, A.L. (2017). Disadvantaged minorities' use of the internet to expand their social networks. *Communication Research, 44,* 467-486. doi: 10.1177/0093650214565925

Günsoy, C., Cross, S.E., Saribau, A., Ökten, I.O., & Kurutaş, M. (2015). Would you post that picture and let your dad see it? Culture, honor, and Facebook. *European Journal of Social Psychology, 45,* 323-335. doi: 10.1002/ejsp.2041

Hales, A.H., Wesselman,, E.D., & Williams, K.D. (2016). Prayer, self-affirmation, and distraction improve recovery from short-term ostracism. *Journal of Experimental Social Psychology, 64,* 8-20. DOI: 10.1016/jesp.2016.01.002

Hales, A. H., & Williams, K. D. (2018). Marginalized individuals and extremism: The role of ostracism in openness to extreme groups. *Journal of Social Issues, 74,* 75-92. DOI: 10.1111/josi.12257

Harinck, F., Shafa, S., Ellemers, N., & Beersma, B. (2013). The good news about honour culture: The preferences for cooperative conflict management in the absence of insults. *Negotiation and Conflict Management Research, 6,* 67-78. DOI: 10.1111/ncmr.12007

Harmon-Jones, C., Bastian, B., & Harmon-Jones, E. (2016). The discrete emotions questionnaire: A new tool for measuring state self-reported emotions. *PLoS ONE, 11,* 1-25. DOI: 10.1371/journal.pone.0159915

Hartgerink, C.H.J., van Beest, I., Wicherts, J.M., & Williams, K.D. (2015). The ordinal effects of ostracism: A meta-analysis of 120 Cyberball studies. *PLOS One, 10*(5), e0127002. doi: 10.1371/journal.pone.0127002

Hashimoto, H., & Yamagishi, T. (2016). Duality of independence and interdependence: An adaptationist perspective. *Asian Journal of Social Psychology, 19*, 286-297. DOI: 10.1111/ajsp.12145

Hashimoto, H., & Yamagishi, T. (2013). Two faces of interdependence: Harmony seeking and rejection avoidance. *Asian Journal of Social Psychology*, *16*, 142-151. DOI: 10.1111/ajsp.12022

Ijzerman, H., & Cohen, D. (2011). Grounding cultural syndromes: Body comportment and values in honor and dignity cultures. *European Journal of Social Psychology*, *41*, 456-467. doi: 10.1002/ejsp.806

James, W. (1950). *The principles of psychology*. New York, NY: Dover.

Johnson, W.B., Barnett, J.E., Elman, N.S., Forrest, L., & Kaslow, N.J. (2013). The competence constellation model: A communitarian approach to support professional competence. *Professional Psychology: Research and Practice, 44*, 343-354. doi: 10.1037/a0033131

Kim, Y.-H., & Cohen, D. (2010). Information, perspective, and judgments about the self in face and dignity cultures. *Personality and Social Psychology Bulletin*, *36*, 537-550. DOI: 10.1177/0146167210362398

Kitayama, S., Mesquita, B., & Karasawa, M. (2006). Cultural affordances and emotional experience: Socially engaging and disengaging emotions in Japan and the United States. *Journal of Personality and Social Psychology, 91*, 890-903. DOI: 10.1037/0022-3514.91.5.890

Konrad, K.A., & Morath, F. (2012). Evolutionarily stable in-group favoritism and out-group spite in intergroup conflict. *Journal of Theoretical Biology, 306*, 61-67. doi: 10.1016/j.jtbi.2012.04.013

Kwak, D., & Kim, W. (2017). Understanding the process of social network evolution: Online-offline integrated analysis of social tie formation. *PLOS One, 12*(5), e0177729. doi: 10.1371/journal.pone.0177729

Kyom. (2016). "The internet isn't real. You can say what you like. How people act on the internet is no relation to their real self." Drama in /r/relationships when one user doesn't judge someone for cyber-bullying. *Reddit*. Retrieved from https://www.reddit.com/r/SubredditDrama/comments/3dve99/the_internet_isnt_real_you_can_say_what_you_like/

Lee, H.I., Leung, A.K.-Y., & Kim, Y.-H. (2014). Unpacking east-west differences in the extent of self-enhancement from the perspective of face versus dignity culture. *Social and Personality Psychology Compass*, *8*, 314-327. DOI: 10.1111/spc3.12112

Leitner, J.B., Hehman, E., Deegan, M.P., & Jones, J.M. (2014). Adaptive disengagement buffers self-esteem from negative social feedback. *Personality and Social Psychology Bulletin*, *40*, 1435-1450. DOI: 10.1177/0146167214549319

Leung, A.K.-Y., & Cohen, D. (2011). Within- and between-culture variation: Individual differences and the cultural logics of honour, face, and dignity cultures. *Journal of Personality and Social Psychology*, *100*, 507-526. DOI: 10.1037/a0022151

Li, Q. (2008). A cross-cultural comparison of adolescents' experience related to cyberbullying. *Educational Research*, *50*, 223-234. DOI: 10.1080/00131880802309333

Liu, J., Huo, Y., Chen, Y., & Song, P. (2018). Dispositional and experimentally primed attachment security reduced cyber aggression after cyber ostracism. *Computers in Human Behaviour*, *84*, 334-341. DOI: 10.1016/j.chb.2018.02.040

Ma, R., Huang, Y-.C., & Shankar, O. (2011). Social networks and opportunity recognition: A cultural comparison between Taiwan and the United States. *Strategic Management Journal*, *32*, 1183-1205. doi: 10.1002/smj.933

Maitner, A.T., Mackie, D.M., Pauketat, J.V.T., & Smith, E.R. (2017). The impact of culture and identity on emotional reactions to insults. *Journal of Cross-Cultural Psychology*, *48*, 892-913, DOI: 10.1177/0022022117701194

Ma, T.-L., & Bellmore, A. (2016). Early adolescents' responses upon witnessing peer victimization: A cross-culture comparison between students in Taiwan and the United States. *International Journal of Developmental Science*, *10*, 33-42. DOI: 10.3233/DEV-150176

Markus, H.R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, *98*, 224-253. DOI: 10.1037/0033-295X.98.2.224

Megías, C.F., Mateos, J.C.P., Ribaudi, J.S., & Fernández-Abascal, E.G. (2011). Validación Española de una batería de películas para inducer emociones. *Psicothema*, *23*, 778-785. Retrieved from http://www.psicothema.es/pdf/3956.pdf

Nisbett, R. E., & Cohen, D. (1996). Culture of honour: The psychology of violence in the South. Denver, CO: Westview Press.

Obaidi, M., Bergh, R., Akrami, N., & Anjum, G. (2019). Group-based relative deprivation explains endorsement of extremism among Western-born Muslims. *Psychological Science*, 1-10. doi: 10.1177/0956797619834879

van Osch, Y., Breugelmans, S.M., Zeelenberg, M., & Bölük, P. (2013). A different kind of honor culture: Family honor and aggression in Turks. *Group Processes & Intergroup Relations*, *16*, 334-344. doi: 10.1177/1368430212467475

O'Sullivan, P.B., & Flanagin, A.J. (2003). Reconceptualizing 'flaming' and other problematic messages. *New Media & Society, 5*, 69-94. DOI: 10.1177/1461444803005001908

Ouwerkerk, J.W., Kerr, N.L., Gallucci, M., & van Lange, P.A.M. (2005). Avoiding the social death penalty: Ostracism and cooperation in social dilemmas. In K.D. Williams, J.P. Forgas, & W. von Hippel (Eds.), *Sydney Symposium of Psychology Psychology Series. The social outcast: Ostracism, social exclusion, rejection, and bullying* (pp. 321-332). New York, NY: Psychology Press.

Pearce, K.E., & Vitak, J. (2016). Performing honor online: The affordances of social media for surveillance and impression management in an honor culture. *New Media & Society, 18*, 2596-2612. doi: 10.1177/1461444815600279

Peng, A.C., & Tjosvold, D. (2011). Social face concerns and conflict avoidance of Chinese employees with their western or Chinese managers. *Human Relations, 64*, 1031-1050. Doi: 10.1177/0018726711400927

Pfundmair, M., Graupmann, V., Frey, D., & Aydin, N. (2015). The different behavioral intentions of collectivists and individualists in response to social exclusion. *Personality and Social Psychology Bulletin, 41*, 363-378. doi: 10.1177/0146167214566186

Phillips, W. (2016). *This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture*. Boston, MA: MIT Press.

Pinto, I.R., Marques, J.M., Levine, J.M., & Abrams, D. (2010). Membership status and subjective group dynamics: Who triggers the black sheep effect? *Journal of Personality and Social Psychology, 99*, 107-119. doi: 10.1037/a0018187

Poon, K.-T., & Chen, Z. (2016). Assuring a sense of growth: A cognitive strategy to weaken the effect of cyber-ostracism on aggression. *Computers in Human Behavior, 57*, 31-37. DOI: 10.1016/j.chb.2015.12.032

Postmes, T., Spears, R., & Lea, M. (1998). Breaching or building social boundaries? SIDE-effects of computer-mediated communication. *Communication Research, 25*, 689-715. doi: 10.1177/009365098025006006

Prati, F., & Giner-Sorolla, R. (2017). Perceiving mixed valence emotions reduces intergroup dehumanization. *Cognition and Emotion, 32*, 1018-1031. doi: 10.1080/02699931.2017.1383885

Rodriguez Mosquera, P.M., Fischer, A.H., Manstead, A.S.R., & Zaalberg, R. (2008). Attack, disapproval, or withdrawal? The role of honour in anger and shame responses to being insulted. *Cognition and Emotions, 22*, 1471-1498. DOI: 10.1080/02699930701822272

Rullo, M., Presaghi, F., & Livi, S. (2015). Reactions to ingroup and outgroup deviants: An experimental group paradigm for black sheep effect. *PLOS One, 10*, e0125605. doi: 10.1371/journal.pone.0125605

Schaafsma, J., & Williams, K.D. (2012). Exclusion, intergroup hostility, and religious fundamentalism. *Journal of Experimental Social Psychology, 48*, 829-837. DOI: 10.1016/j.jesp.2012.02.015

de Seta, G. (2013). Spraying, fishing, looking for trouble: The Chinese internet and a critical perspective on the concept of trolling. *The Fibreculture Journal, 22,* 301-318. Retrieved from http://fibreculturejournal.org/wp-content/pdfs/FCJ-167Gabriele%20de%20Seta.pdf

Severance, L., Bui-Wrzosinska, L., Gelfand, M.J., Lyons, S., Nowak, A., Borkowski, W., … & Yamagushi, S. (2013). The psychological structure of aggression across cultures. *Journal of Organization Behavior, 34,* 835-865. DOI: 10.1177/0146167214549319

Smith, P.B., Easterbrook, M.J., Blount, J., Koc, Y., Harb, C., Torres, C., … & Rizwan, M. (2017). Culture as perceived context: An exploration of the distinction between dignity, face and honor cultures. *Acta de Investigación Psicológica, 7,* 2568-2576. doi: 10.1016/j.aipprr.2017.03.001

Synnott, J., Coulias, A., & Ioannou, M. (2017). Online trolling: The case of Madeleine McCann. *Computers in Human Behavior, 71,* 70-78. doi: 10.1016/j.chb.2017.01.053

Thacker, S., & Griffiths, M. D. (2012). An exploratory study of trolling in online video gaming. *International Journal of Cyber Behaviour, Psychology and Learning, 2,* 17-33. DOI: 10.4018/ijcbpl.2012100102

Ting-Toomey, S., Gao, G., Trubisky, P., Yang, Z., Kim, H.S., Lin, S.-L., & Nishida, T. (1991). Culture, face maintenance, and styles of handling interpersonal conflict: A study in five cultures. *The International Journal of Conflict Management, 2,* 275-296. Retrieved from https://www.researchgate.net/publication/235266557_Culture_Face_Maintenance_and_Styles_of_Handling_Interpersonal_Conflict_A_Study_in_Five_Cultures

Uskul, A.K., & Cross, S.E. (2018). The social and cultural psychology of honour: What have we learned from researching honour in Turkey? *European Review of Social Psychology, 30,* 39-73. DOI: 10.1080/10463283.2018.1542903

Uskul, A.K., & Over, H. (2014). Responses to social exclusion in cultural context: Evidence from farming and herding communities. *Journal of Personality and Social Psychology, 106,* 752-771. DOI: 10.1037/a0035810

Verdery, A.M., Mouw, T., Edulblute, H., & Chavez, S. (2018). Communication flows and the durability of a transnational social field. *Social Networks, 53,* 57-71. doi: 10.1016/j.socnet.2017.03.002

Vignoles, V.L., Owe, E., Becker, M., Smith, P.B., Easterbroook, M.J., Brown, R. … Brambilla, M. (2016). Beyond the 'East-West' dichotomy: Global variation in cultural models of selfhood. *Journal of Experimental Psychology: General, 145,* 966-1000. DOI: 10.1037/xge0000175

Wang, L., Zheng, J., Meng, L., Lu, Q., & Ma, Q. (2016). Ingroup favoritism or the black sheep effect: Perceived intentions modulate subjective responses to aggressive interactions. *Neuroscience Research, 108,* 46-54. doi: 10.1016/j.neures.2016.01.011

Williams, K.D. (2009). Ostracism: A temporal need-threat model. In M.P. Zanna (Ed.), *Advances in experimental social psychology, Vol. 41* (pp. 275-314). San Diego, CA: Elsevier Academic Press.

Williams, K.D., Cheung, C.K.T., & Choi, W. (2000). Cyberostracism: Effects of being ignored over the internet. *Journal of Personality and Social Psychology, 79,* 748-762. DOI: I0.1037//O022-3514.79.5.748

Yilmaz, T., & Bekaroğlu, E. (2020). Ayrık duygular ölçeğinin Türkçeye uyarlama, güvenirlik ve geçerlik çalışması. *Turkish Studie – Social, 15,* 2233-2244. doi: 10.29228/TurkishStudies.40502

Zadro, L., Williams, K.D., & Richardson, R. (2004). How low can you go? Ostracism by a computer is sufficient to lower self-reported levels of belonging, control, self-esteem, and meaningful existence. *Journal of Experimental Social Psychology*, *40*, 560-567. DOI: 10.1016/j.jesp.2003.1

Chapter 5: A Bystander State of Mind

Bystander Reactions to an Experimental In-Game Trolling Situation

Previous research has suggested that reporting trolls to an authority figure – the developer, in the case of online games – is the most effective trolling deterrent currently available to bystanders, but also the least-used recourse. We do not yet know, however, what motivates a bystander to report a troll, or which personal characteristics can influence that decision. An experiment was thus conducted to determine which factors during the game and which individual differences influence bystanders' intention to report a troll. A naïve participant was placed in an online team game (League of Legends) with two confederates who began to troll one another – a non-critical, but unpleasant, situation. The intention to report, as well as individual differences, were assessed via a post experiment questionnaire. Results showed that trolls were only reported if they scored low on warmth or if their trolling disrupted the cohesion of the team; emotion and perceived competence of teammates had no significant impact on bystanders' intention to report the troll, although the qualitative portion of our data would suggest that perceived competence may have an effect on the decision to report with a different, more competitive sample of League of Legends players. Finally, we suggest future research into trolling interventions.

At the time of submission, this chapter was under review as:

Cook, C.L., Antheunis, M.L., & Schaafsma, J. (2020). Not my job: Bystander reactions to an experimental in-game trolling situation. *Communication Monographs*. Manuscript under review.

Introduction

Despite its very recent explosion into our leisure time, the concept of an online game is not new. Before the current triple-A gaming giants like League of Legends, Overwatch, and World of Warcraft, the internet was populated with simple text-based online worlds called either MUDs (Multiple User Dungeon) or MOOs (an object-oriented MUD; Carroll et al., 2001; Dourish, 1998). In these 'spatial environments' (Erickson, 1993), players would create characters they would play in this fictional online world and forge relationships with other characters played by strangers (Dibbell, 1993; Jacobson, 2001). It was in one of these early worlds, LambdaMOO, that one of the first recorded incidents of severe online harassment took place (Dibbell, 1993). Dibbell (1993) describes the situation as being a "rape in cyberspace". One player – Mr. Bungle – forced several other players in the MOO to perform violent sexual acts against themselves and each other. Eventually, when they had enough of their self-appointed despot, the rest of the community banded together to summon a powerful wizard-like character who could banish Mr. Bungle's avatar from the server, thus ending his tyranny. In essence, when threatened, the players came together to bring justice to their virtual world.

Nowadays, justice in online games is still an ever-present issue (see Cook et al., 2018). Gamers today would probably call people like Mr. Bungle 'trolls', and his behaviour 'trolling', which is the use of the gameplay or chat mechanics of a game at another player's expense (see Cook et al., 2019). The people who banded together to help the victims in Dibbel's (1993) account were bystanders. Contrary to the findings in traditional bystander studies though, which predict low rates of intervention among bystanders in larger groups (see Darley & Latané, 1968), the bystanders to Dibbel's (1993) trolling situation came together in aid of the victims. More recent studies specific to online gaming have found that bystanders do still respond to trolls in

this context, but their preferred methods are less effective than those of Dibbell's (1993) bystanders (Cook et al., 2018). The most effective method of deterring trolling available to today's players is reporting trolls to game developers – a recourse that was not always available in Dibbell's (1993) day – but instead, current bystanders generally join in the trolling or try to attack the troll instead (Cook et al., 2018). There appears to be something about either gaming or gamers that makes bystanders more active in trolling situations than bystanders in other contexts, but as of yet, we do not know what makes these bystanders make the decisions that they do, be that to intervene on the victim's behalf or join in the trolling.

The specific type of intervention of particular interest to academia and society alike is reporting, as increased reporting should theoretically reduce trolling levels overall (Cook et al., 2018), but even among bystanders who report, there is still room for many factors to affect the relationship between trolling and reporting. Fox, Gilbert and Tang (2018), in one of the few studies directly addressing reporting behaviour, found that there are differences in terms of the gender of the player, with women noting and reporting more sexual language as offensive and men, as well as a difference between senior and novice players' reporting choices. Research has also shown that the actions and speech of others online can have a massive impact in cyberbullying and trolling situations (e.g., Allison & Bussey, 2016; Brody & Vangelisti, 2016; Cook et al., 2019). In addition, there is evidence that a bystander's internal state could affect their reporting behaviour (e.g., Henik, 2015). While these factors have been examined quite extensively in trolling literature (e.g., Cheng et al., 2017; Cook et al., 2018; 2019), they are only a small part of the literature on reporting behaviour (e.g., Gollwitzer & van Prooijen, 2017; Henik, 2015). Thus, we know that a person's internal state, as well as the person's fellow

players, are important in trolling circumstances; what we do not yet know is how they affect the relationship between the presence of trolling and the decision to report a troll.

The present study aims to explore the antecedents of reporting behaviour in the online gaming context and determine how a) the victim's reaction could moderate, and b) the bystander's emotions and c) their perceptions of the troll and victim, particularly their competence, could mediate the relationship between trolling and reporting. To do this, an experiment will be conducted in which naïve gamers are put into an experimental game of League of Legends (Riot Games, 2009), where they compete against a team of three computer opponents. Their prior experience with the game will be evaluated to control for any effect game experience may have. The two other members of their team are confederates, and one will begin to troll the other according to a predesigned script. Through this experience, players should have some kind of emotional reaction to the trolling they are witnessing, and they should develop opinions about the competence of the two confederates (the troll and the victim). They will also work in a team with these strangers, and this could be either a cohesive or a discordant experience. All of these factors could impact the bystanders' reporting decisions. However, the victim may or may not respond to the in-game trolling, giving the bystander more or less information to use in their decision to report and theoretically moderating the relationship between trolling and reporting. At the game's end, participants will be asked anonymously whether they would have reported (using the game's reporting function) a member of their team, and why or why not. If we can understand the conditions required to elicit a report from a bystander, we can learn how best to bring justice to the virtual worlds we inhabit today.

Theoretical background

Within trolling literature, the act of reporting trolls has received only limited attention (Cook et al., 2018; Cook et al., 2019; Fox et al., 2018). It is, however, a major component of the game industry's fight against the most negative forms of trolling, referred to commonly as online toxicity (see Blackburn & Kwak, 2014; Kwak et al., 2015). Reporting is the method players can use to alert game developers to punishable behaviour, according to the guidelines published by the company in question (for an example, see Riot Games, 2020). Typically, it happens at the end of a game, although certain titles do allow reporting during the game (see Blizzard, 2004). Academics have often noted that reporting is only one of many possible reactions to trolling (e.g., Herring et al., 2002), but more recent work has suggested that trolls are effectively warded off from future trolling by being reported and subsequently punished (Cook et al., 2018).

Nevertheless, reporting is not without its risks and problems. Although an effective trolling deterrent (Cook et al., 2018), research has also found that players will sometimes abuse the report button at the end of a game, using it as a way to get revenge on other players who might have performed poorly or criticized them during the game, or even as a way to troll other players (Balci & Salah, 2015; Fox et al., 2018). Although Balci and Salah's (2015) findings when examining player reports would suggest that the falsely-reported far outnumber the justly-reported, the game they studied was more akin to a digitalized board game than a competitive multiplayer video game like the one used in the present study. In short, we do not yet know how generalizable these results are to the rest of the world of online gaming. To put these results to the test, participants in the present study will be placed in one of two types of conditions: with a troll present, or without a troll present. In this way, we can answer the following question:

RQ1: How often do gamers report players when no obvious offense has been committed?

However, the mere presence of trolling is not likely to cause a report on its own; if that were the case, all trolls would be reported all the time, so there are almost certainly other factors that influence this decision. In Valentine and Godkin's (2019) study of whistleblowers – a phenomenon akin to reporting in an online game – social consensus on an entity's deservedness of punishment was just as strong a predictor of whistleblowing as the perceived seriousness of the offense. Translated to an online gaming situation, if no one else on the team responds to the trolling, the player might believe that the behaviour is being tolerated, and therefore may choose not to make a report of the perceived perpetrator. However, if the team does respond, as long as they do not respond positively to the trolling, the player may feel more secure in reporting the offense. Hence, how the group responds to trolling should also help determine whether or not a player engages in reporting (McGlynn & Richardson, 2014; Misch et al., 2018; Waytz et al., 2013). We can see this reflected in Bastiaensens and colleagues' (2014) study of cyberbullying bystanders, in which bystanders expressed the intent to conform to other bystanders' behaviours when faced with a cyberbullying situation. Slonje, Smith, and Frisén (2012) also report that talking about what happened or is happening is an effective intervention strategy when it comes to cyberbullying incidents.

In the present study, participants playing the game (League of Legends) are on teams of three. In the conditions in which a troll is present, this means that the team consists of a troll, the troll's victim, and our participant (the bystander). When a troll is present, how the victim reacts to the trolling is what establishes a contrasting social norm (as opposed to the troll's norm, which is trolling) the participant can follow to achieve social consensus within the team. It would not be practical to test every possible reaction a victim could have in a single study, as the list of options is extensive (see Cook et al., 2018, for examples). Thus, we have reduced the options to

two: a victim who responds neutrally and a victim who does not respond at all. In this way, we can determine how the basic idea of another teammate responding to trolling influences the decision to report the troll. Given results from earlier studies of whistleblowing and trolling, we anticipate the following:

H1) If the victim sends messages in-game *while being trolled*, the bystander will be more likely to report the troll than if the victim remains silent.

Moral Emotions and Reporting as Justice Behaviour

Literature on justice behaviours like reporting would suggest that there are other variables that can also influence the decision to engage in reporting or not. One of the most-researched of these are particular emotions, called "moral emotions" (Ellemers et al., 2019). Different moral emotions precede different types of justice behaviours, namely punishment or reward for the perpetrator or victim. In the present study, we are focusing specifically on reporting behaviour, which is an indirect form of punishment focused on the perpetrator: the troll. According to Gollwitzer and van Prooijen (2016), the act of bringing justice via punishment is typically motivated by anger or moral outrage, meaning that the justice-bringers are extremely upset by the perpetrator's action and want to make it right. As previously mentioned, the closest real-world analogue to in-game reporting is whistleblowing (see Henik, 2015). Many studies have found that whistleblowers, too, are frequently motivated by the anger they feel at witnessing the injustice they plan to make public (Cassematis & Wortley, 2013; Henik, 2015; McGlynn & Richardson, 2014), but there are several important differences between this context and that of an online game that could alter this relationship between anger and justice. The most important of these is the near guarantee of anonymity that a gamer has when reporting another player. This anonymity essentially removes the fear of retaliation from colleagues of teammates that normally

comes with reporting behaviours like whistleblowing (see Henik, 2015). As such, we need to consider how the gaming medium, and particularly its accompanying anonymity, may encourage or discourage the relationship between moral outrage (anger) and justice behaviour (reporting).

There are several existing theories that suggest that this anger will be intensified due to the anonymity inherent to online gaming. One such theory is the social identity model of deindividuation effects (SIDE), which posits that anonymous people in groups tend to polarize in terms of their attitudes and feelings (Postmes et al., 1998). Previous research has suggested that this model applies in the gaming context, particularly when people are on teams (e.g., Cook et al., 2019), leading to an intensification of the emotional experience when they witnessed or were a victim of trolling (see also Cook et al., 2018). Therefore, if participants are angered by witnessing trolling in the present experiment, this emotion should theoretically be intensified by the online, anonymous context.

This is also in line with both moral psychology (Gollwitzer & van Prooijen, 2016) and our analogue reporting experience, whistleblowing (Cassematis & Wortley, 2013; Henik, 2015; McGlynn & Richardson, 2014), suggesting that reporting a troll is like another form of whistleblowing, and that the mechanisms behind both may be similar. Other gaming-specific research has also suggested that the reporting function in games can be used as a form of either trolling in and of itself (e.g., Fox et al., 2018), or a form of retaliation against trolls (e.g., Cook et al., 2018), both of which are also preceded by anger, often called ‘tilt’ in the gaming context. Thus, even if they are not specifically motivated by justice, but rather by revenge, the relationship between moral outrage/anger should still hold in the present study. That said, this negative emotional amplification should only be the case when trolling is present, as without the trolling behaviour, there should be no trigger for the moral emotion. Since this emotion is

directed at the troll, and our confederate victim is not actively trying to garner sympathy in the present study, there is no evidence to suggest that the victim's degree of agency will affect the emotional experience of bystanders. We thus expect the following:

> H2) Bystanders are likely to experience negative emotions like anger when witnessing a victim being trolled, irrespective of the victim's reaction, which should consequently raise the likelihood that the bystander will report the troll.

The Principle of Deservedness and Reporting Behaviours

In addition to moral emotions, it is critical to note that bystanders do not only have themselves to consider when they make the decision to engage in justice or revenge behaviours. Another factor that is key when it comes to serving justice is the bystanders' evaluation of the other actors involved in the trolling situation. Essentially, to elicit a justice behaviour, the bystander should believe that the perpetrator deserves punishment and the victim deserves support in order to commit to reporting the troll. In moral psychology literature, this is called the principle of deservedness (Hagai & Crosby, 2016). If mitigating circumstances are present, such as reciprocal norm violations from a victim, a bystander could potentially decide that the original perpetrator does not deserve a harsh punishment, or that both deserve equal punishment (see Baumert & Schmitt, 2016, for further discussion). In a gaming situation, this could apply to trolling victims who retaliate against their aggressor; if bystanders believe that victim and perpetrator are equally wrong, they may not report the original troll, or may choose to report both the troll for initiating the interaction and the victim for continuing it. Being angry at a person is not necessarily enough on its own to induce reporting; the person also has to believe the perpetrator deserves the punishment, and that the victim deserves justice.

However, contrary to most bystander situations, which are typically performed in person (e.g., Abbott & Cameron, 2014; Darley & Latané, 1968; Fischer et al., 2011), the online context in the present study prevents normal cues like facial expressions or physical mannerisms from being identified. In addition, all players use pre-created anonymous accounts in the present study to prevent players from evaluating players on linguistic cues that could appear in their personal usernames (Harari et al., 2015). Because of these experimental limitations, there are only two ways in which our participants can judge the deservedness of their teammates: 1) the confederates' gameplay, and 2) their chat messages. From this they can determine how strong a player the person in question is, and how friendly or unfriendly they are, but little beyond these two variables, which map onto Fiske and colleagues' (2002) dimensions of personality: competence and warmth. If a person is very warm, they are someone who is friendly and compassionate toward others, while a person who is highly competent is someone who commands respect from others, whether they actively seek the respect or not (Fiske et al., 2002). The higher a person scores on these two qualities, generally, the better they are received by others.

Translated to an online gaming situation, we anticipate the perceived competence of the victim and bystander being particularly important in terms of impact on reporting intentions. This expectation is based on two main ideas: 1) the principles of deservedness and equity as described in moral psychology (Baumert & Schmitt, 2016; van den Bos & Bal, 2016; Montada & Maes, 2016), and 2) online games like League of Legends serve as a competitive setting. We have already seen the principle of deservedness (Hagai & Crosby, 2016), but put specifically into the context of a mobile online battle arena (MOBA) like League of Legends, this principle would dictate that the more a player contributes toward the goals of the team, the more deserving they

are of reward, and the less deserving they are of punishment. If a player inhibits the attainment of the team's goals, as trolls are often purported to do (Cook et al., 2018), the more deserving they are of punishment, and the less deserving of rewards. It is possible that bystanders could interpret the troll's trolling as in-game engagement, but earlier studies would suggest that this is the rarer option (see Cook et al., 2018; 2019); more often than not, trolling should be perceived as inhibiting the goal of winning the game, as they are taking away the attention of the victimized team member. As such, trolling should theoretically reduce trolls' perceived competence, thereby increasing the chance for a report at game's end. However, given the fact that in our agent victim conditions, the victim is talking about the game in question, this is also likely to have an impact on their perceived competence. They are more actively engaging with the game, which should boost perceived competence, and consequently lead to more reports from our participants as well.

The second principle, equity, states essentially that in the absence of extenuating circumstances (as dictated by the principle of deservedness, for instance), all people should be treated equally (see Baumert & Schmitt, 2016, for a full discussion of equity theory). In the gaming context, a popular or well-known player who trolls should be punished in the same way and to the same degree as a casual player who trolls, just as they should be praised equally for impressive displays of skill. This makes competence all the more powerful, as it determines the deservedness of punishment or reward when all else is equal. In the absence of trolling, competence should not have any relationship with reporting, as there is nothing meriting punishment to begin with. We thus propose the following hypotheses:

H3a) When trolling occurs, bystanders are likely to give a lower competence evaluation to trolls, irrespective of the victim's reaction, and are thereby more likely to report the troll at game's end.

H3b) When trolling occurs, if the victim speaks up, the bystander is likely to give a higher competence evaluation to the victim, and will consequently be more likely to report the troll at game's end.

Perceived Group Cohesion as a Predictor of Reporting

We have established that the bystander's personal emotions at the time, as well as their evaluation of the other players, have a role to play in determining whether or not the bystander engages in justice behaviours or not. However, at its core, the social dynamic *between* players on a team is also likely to affect a player's decision to report a teammate. Previous studies have established that teamwork is a critical factor that contributes to the win or loss of a team in-game (see Mora-Cantallops & Sicilia, 2019), and loss-streaks have been mentioned repeatedly by trolls as being a catalyst to trolling behaviour (Cook et al., 2018). As such, we would expect a positive team atmosphere and high cohesion to lower the chances of a player getting reported, as a higher degree of team cohesion would increase the chances of a win and decrease the chances of a loss. With low cohesion and poor teamwork, the chances of seeing a player reported would increase, as the chance of a win diminishes and the chance of further trolling increases the worse a team works together toward their goal. In short, the more the victim or troll risks the bystander's personal goals, the more likely they are to be reported.

That said, extant literature adds an additional layer to the decision of whether to report someone: the potential reporter's inner conflict between their loyalty to their team and their sense of fairness (Misch et al., 2018; Waytz et al., 2013). Loyalty to the group reduces the chance of

reporting, while a strong sense of fairness increases it. In the present context of a standard match-made game of League of Legends, there are no pre-existing ties between players (see Alman & McKay, 2017 for a complete discussion of matchmaking in online games), and so loyalty to the group should be determined by how the group acts in-game (see Kahn & Williams, 2016). If the players work together and avoid getting in one another's way, it is in players' best interests to remain loyal to the team, as this will increase the likelihood of achieving their personal goal: victory (Mora-Cantallops & Sicilia, 2018a). Since our conditions that include victim agency are likely to increase the victim's apparent engagement with the game, this will probably increase the perceived group cohesion as well. If the bystander were to report a member of this cohesive team, they are lessening their chances of being paired with them again in the future, thus potentially decreasing their chances of future victories as a matter of consequence. However, in the present study, there are no pre-existing ties between players, and in the trolling conditions, the players are not building any ties through their behaviour. Thus, with no loyalty at risk, the bystander is theoretically free to report any of them (Kahn & Williams, 2016; Lee et al., 2019; Mora-Cantallops & Sicilia, 2018a). It would actually be beneficial to do so, as it would reduce the chance of being paired with people who work together poorly in future games. We thus predict the following:

> H4) When trolling occurs, bystanders are likely to perceive a lower level of group cohesion than if no trolling occurs, especially if the victim speaks up. This reduction of perceived group cohesion should result in an increased likelihood for the bystander to report the troll at game's end.

Method

Participants and Design

Participants were 97 gamers (88 men and 9 women) between the ages of 18 and 32 (M = 22.45, SD = 3.10) who had at least some prior experience playing League of Legends. This generally fits the demographics of most League of Legends players, who are mostly young – 85% of the player base is between the ages of 16 and 30 – and male, 90% of players (Gallegos, 2012). These gamers were recruited via posted advertisements on various gaming forums across Europe and the United States, as well as social networking sites. Participants were predominantly Dutch (42.3%); the rest either came from other countries in Europe (40.2%), North America (11.3%), or Africa (4.1%). Only two participants did not disclose their nationality. Most of our participants had completed either their upper secondary education (high school or local equivalent) or the equivalent of a Bachelor's degree at the time of the experiment (81.4%), while the rest had either a basic education (pre-high school; 7.2%) or a graduate degree (11.3%). Due to our international sample and the experiment being in English, we also asked them to disclose their estimated English proficiency. The majority professed either an advanced (30.9%) understanding or fluency (36.1%) in the language, and there were several native speakers (18.6%). The rest (14.4%) claimed either a basic (1 participant) or intermediate command of English. In terms of their League of Legends experience, the vast majority had been playing the game for a year or more (90.7%), and played the game on a weekly basis at least (67%). Participants themselves were paid five euros via PayPal for their participation.

The experiment itself took a 2 (no trolling versus trolling) x 2 (victim remains silent versus victim speaks) format, and participants were randomly assigned one of these four possible

conditions upon their scheduling in Doodle, an online scheduling website. Scripts for all conditions are found in Appendix A.

Procedure

From the advertisements, potential participants were directed to a Doodle page where they received a more detailed description of the study. The study was presented as a study of teamwork among strangers in an online setting; its true nature as an examination of trolling and bystander behaviour was kept secret until the debriefing. Potential participants were then offered the chance to select a time slot in which to participate.

An hour before their chosen time slot, participants were sent an e-mail with a link to the Qualtrics pre-test. After filling out an electronic consent form, participants answered some demographic questions (i.e., age, gender, education level, nationality, English proficiency), followed by questions regarding their experience with League of Legends specifically (ex., "How often do you play League of Legends?", "Do you have a competitive League of Legends account?"), as well as any other online multiplayer game experience. The pre-test concluded with a questionnaire asking participants to detail their emotional experience at that moment.

At this point, a screen appeared instructing the participant to log into League of Legends using the credentials provided. A code given by the researcher after the game was required to move past this screen and onto the post experiment questionnaire, preventing participants from simply skipping the game portion of the study. It should be noted that it was our initial intention to use the in-game chat and statistics as additional indicators of competence and emotion, but due to an unexpected update, our game data was irretrievable from Riot Games' servers. Thus, the only information we have from the game itself is approximate duration. Most games were within

the same amount of time (approximately 20 minutes), so this was not used in our analyses. However, participants did use an unaltered form of the game client and game in the experiment.

The custom game took place on the 3v3 map 'Twisted Treeline' with two confederates who chatted according to script throughout the game. Every participant won their game, and games lasted a maximum of 30 minutes. In terms of gameplay, the victim either spoke or remained silent and the troll either trolled (operationalized as flaming) or remained silent, depending on the experimental condition of the participant (all scripts are available in Appendix A), and both played the game with the goal of finishing as quickly as possible. This was done to both limit participant fatigue and provide a realistic simulation of other participants playing to win. Due to League of Legends' champion rotation system, which changed the free-to-play champions available every week, participants and confederates were not able to play the same champion (character) every game. However, given the link between trolling and character selection (Burkholder, 2019; Lee et al., 2019) the troll confederate always played a melee-style champion (a character who attacks up close) and the victim always played a ranged-style champion (a character who attacks from afar) to keep reasonable consistency between games. Confederates also attempted to keep their play consistent to avoid any effect of poor or excellent play, as this too could potentially bias participants. All human players were on the same team against three AI-controlled opponents set to 'beginner', the easiest level. A researcher 'spectated' the game – a mode in which players are not directly involved in the game, but can watch all the action as if they were playing – all whilst video recording the game on their own computer. All participants were made aware that this recording would take place and consented to the recording. At the end of the game, the researcher sent a message to all three players (one naïve, two confederates) with the code to enter into the Qualtrics page that would allow them to

continue the study. The participant then exited the game and returned to Qualtrics to enter the code, thus beginning the post experiment questionnaire.

During the post experiment questionnaire, participants answered the aforementioned open questions regarding their in-game experience (ex., “How did you feel during the game? Describe your experience”). They then completed a second set of questions asking about their emotional experience during the game, a group climate measure, and interpersonal evaluations of each confederate. These completed, the participants were presented with a debriefing text, explaining the true nature of the study. They were also requested not to talk about the study until the data collection phase was completed. Finally, they were offered the chance to fill out a payment form to resubmit to the researchers to receive their five-euro payment.

Materials

The pre and post experiment questionnaires were administered using Qualtrics, and experimental sessions were arranged using Doodle. The experimental session took place within the League of Legends client. In all instances in which multiple measures are presented, their order of presentation was randomized electronically. Individual items were also randomized in the same way. All measures can be found in Appendix B, and all means and standard deviations of possible mediators are presented in Table 1.

Intent to Report. To determine whether participants would have reported one of our confederates or not, we asked them “If this hadn’t been part of a study, would you have reported any of the other players in-game?” at the end of the post experiment questionnaire. This was an open-ended question for participants, but responses were coded dichotomously – yes = 1, no = 0 – while the explanations were retained separately. The only participants who reported (31 participants) were in trolling conditions.

Table 1.

Descriptive statistics for all scales across all experimental conditions.

Trolling Presence	**Victim Response**	**PANAS (pre-test)**				**PANAS (post-test)**				**Troll Evaluation***						**Victim Evaluation***						**Group Cohesion***	
		Positive		**Negative**		**Positive**		**Negative**		**Warmth**		**Comp.**		**Overall**		**Warmth**		**Comp.**		**Overall**			
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
None	None	2.97	0.66	1.67	0.54	3.04	0.75	1.42	0.45	3.47	0.72	3.29	0.94	0.38	0.80	3.35	0.79	3.26	0.90	3.35	0.74	4.00	0.56
	Speech	3.09	0.71	1.64	0.73	3.07	0.82	1.54	0.67	3.70	0.67	3.22	0.69	3.46	0.61	3.73	0.71	3.26	0.81	3.53	0.65	4.09	0.34
Flaming	None	3.23	0.55	1.67	0.53	2.83	0.75	1.67	0.48	2.51	1.02	2.81	0.67	2.65	0.80	3.24	0.56	3.08	0.63	3.19	0.47	2.73	0.63
	Speech	3.30	0.63	1.44	0.51	3.00	0.87	1.60	0.72	2.29	1.04	2.90	0.63	2.63	0.74	3.57	0.89	3.22	0.66	3.41	0.63	2.53	0.52

Note. * = a higher score denotes a more positive perception, while a lower score denotes a more negative perception. Comp. = Competence.

Emotional Experience. To evaluate participants' emotional experience of the game, we employed the PANAS mood inventory (Watson et al., 1988). The PANAS consists of 20 items referring to different emotions, which users ranked on a from 1 (*not at all*) to 5 (*extremely*). These items are divided into two separate subscales: positive and negative affect. When scored, the total gives an indication of how positive or negative a person was feeling at any time indicated by the researcher in the instructions. The first time participants saw this scale, this referred to the moment of testing, meaning when they filled out the questionnaire before the experimental game of League of Legends. When assessed with our sample, both the positive (M = 3.15, SD = 0.64, α = .87) and negative (M = 1.60, SD = 0.59, α = .87) subscales, as well as the full scale (M = 2.38, SD = 0.47, α = .86), demonstrated sufficient reliability. They completed it a second time after the game. The PANAS received similar reliability scores (positive subscale, M = 2.98, SD = 0.79, α = .90; negative subscale, M = 1.56, SD = 0.59, α = .87; M = 2.27, SD = 0.54, full scale, α = .87) in this second instance.

Warmth and Competence. To evaluate the victim and troll's warmth and competence, we created a scale based upon Fiske's Stereotype Content Model (Fiske et al., 2002). Four keywords representing competence (intelligent, creative, knowledgeable, incompetent) and four keywords representing warmth (friendly, kind, mean, trustworthy) were taken from Fiske and colleagues' (2002) article to make the two subscales of our warmth and competence scale. Participants were asked "Based on the game we just played, I think that Player 1 [or 2], is …" twice – once to rate the troll confederate's warmth and competence, and once to rate the victim confederate's warmth and competence. Participants rated the troll and then the victim on each of the 8 items using a scale from 1 (*strongly disagree*) to 5 (*strongly agree*). When used to evaluate the troll confederate, both the warmth (M = 2.99, SD = 1.06, α = .89) and competence (M = 3.05,

SD = 0.75, α = .76) subscales, along with the full scale (*M* = 3.03, *SD* = 0.83, α = .89), demonstrated acceptable reliability. The same was true of the scale when used to evaluate the victim confederate (warmth, *M* = 3.47, *SD* = 0.75, α = .81; full scale, *M* = 3.37, *SD* = 0.63, α = .83), although the competence subscale was noticeably weaker than the others (*M* = 3.26, *SD* = 0.66, α = .67).

Group Cohesion. The group cohesion measure was an adapted form of Mackenzie's (1983) Group Climate Measure. The original scale comprises twelve statements describing group engagement, conflict, and avoidance behaviour. These were evaluated by participants using a seven-point scale ranging from 0 (*not at all*) to 6 (*extremely*), as in the original scale. To better suit the fact that our participants were strangers, the scale was reduced to the two sub-scales most likely to apply to a trolling situation: conflict and avoidance. To ensure proper measurement of separate constructs, certain items that contained multiple qualifiers (ex: The members *rejected* [qualifier 1] and *distrusted* [qualifier 2] each other [sic.]) were separated into multiple items, and the scale was reduced from seven to five points to better match the other scales in the questionnaire (1 = *not at all*, 5 = *extremely*). This left us with an eight-item scale measuring the negative aspects of group perception. Our adapted version (*M* = 3.32, *SD* = 0.88) achieved good reliability (α = .88) to be used in further analyses and is presented in Appendix B along with all other scales used.

Gaming Experience. To control for participants' experience with League of Legends, we asked them how long they had spent as a player, how often they still played, and if their personal account had a competitive ranking.

Results

Analytical Strategy

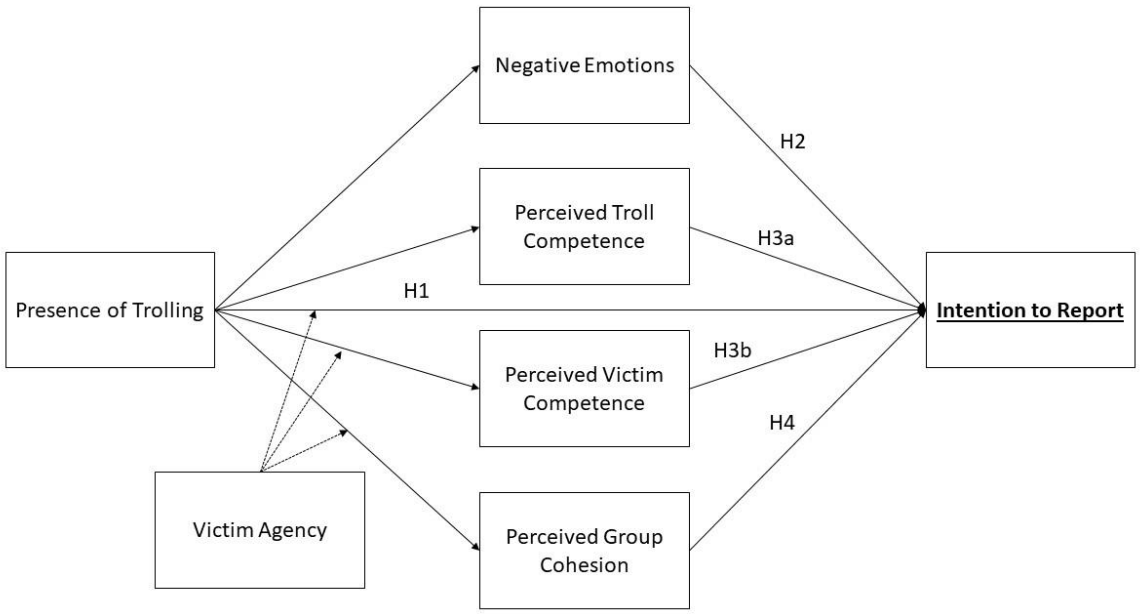
We first checked that all possible predictors that we measured were independent from one another using the "vif" command in the "car" package (Fox & Weisberg, 2019). Once we had determined that there were no violations (all VIFs < 4.46), we tested the linearity of the logit by running a logistic regression to predict reporting behaviour that included interaction terms between each predictor and their log transformation. All interaction terms were insignificant (all *p*s > .07), meaning the linearity of the logit was never violated. However, we encountered a case of complete separation in two instances: our presence of trolling variable, as participants in the non-trolling conditions never expressed an intention to report a player (see Field et al., 2012 for a full discussion of complete separation), and our troll warmth variable, meaning that if the troll's perceived warmth was low enough, they would be always be reported. This answered our first research question (RQ1): Only actual trolls were being reported. The fact that the perceived warmth of the troll also achieved complete separation, with only the least-warmly perceived trolls getting reported, supports the general conclusion of extant literature that trolling is negatively-perceived (e.g., Cook et al., 2018).

We then re-ran our tests of independence and linearity of the logit using only the experimental sessions involving trolling and removed troll warmth as a predictor. This means that, fundamentally, everything we tested is actually an interaction with trolling presence; all analyses were performed only on sessions in which a troll was present. This was to avoid noise in the data. Again, all VIF values were less than 3.68, and there were no violations of the linearity of the logit (*p*s > .07), allowing us to test our hypotheses, presented in Figure 1. To test

our hypotheses, we decided to perform a mediated logistic regression using Hayes' (2017) PROCESS macro. We initially intended to conduct this analysis using Model 8 in this macro, but

Figure 1.

Initial conceptual model.

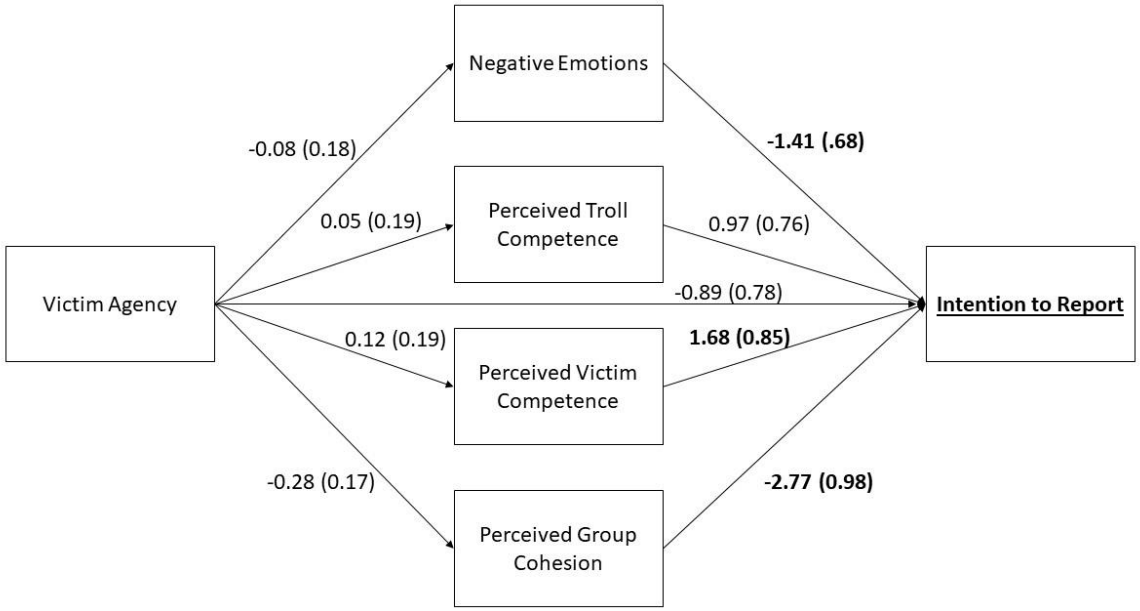


Note. This is the conceptual model of the study's key variables and hypotheses.

we knew at this point that the proposed interaction between the presence of trolling and victim agency (H1) could not be examined, as there was no reporting variance in conditions that did not include trolling. Thus, we were forced to change the model we intended to test, as two of our hypotheses depended on the interaction term that does not exist. We instead decided to test the model presented in Figure 2.

Figure 2.

Redesigned conceptual model and results.



Note. This is the model that we tested in the end, revised from Figure 1 due to the lack of variance in the no-trolling conditions. In this model, trolling presence is removed (due to complete separation), and replaced with victim agency as our last remaining independent variable. Bolded numbers are significant at .05 (all other *p*'s > .07), and numbers in brackets are the coefficient's associated standard error. The full model's final log-order statistics are as follows: -2LL(5) = 46.76, *p* = .01, McFadden = .24, Cox-Snell = .27, Nagelkirk = .37.

In this model – which is Model 4 in the PROCESS macro, with 5000 bootstrap samples – negative emotions, the perceived competence of both the troll and the victim, and the perceived group cohesion all still serve as mediators. However, unlike in our initial model, victim agency is the primary manipulation as opposed to being a moderator. We also included our game experience variables (time spent as a League of Legends player, frequency of League of Legends play, and possession of a competitive rank) as covariates to control for the different levels of experience between our participants. However, none of these covariates were significant, and

thus we reran the analysis without their inclusion. It is these results that we present in the present article. When estimating our model, nonconvergence occurred during bootstrapping, and so confidence intervals should be interpreted with caution.

Testing Our (New) Conceptual Model

Our final results are presented in Figure 2. As the intention to report the troll was a dichotomous variable, all results are expressed in a log-odds metric. As we can see, not only was the direct effect of victim agency on reporting behaviour insignificant (-0.60 [-2.43, 6.87]), but victim agency was not significantly related to any of the mediators (*p*s > .11). This means that H1 – that an agent victim would lead to a higher intention to report among bystanders – is rejected. Although there were no indirect effects between victim agency and the intention to report, some mediators did affect the intention to report. Negative emotions (-1.41 [-2.75, -0.08]) had a negative relationship with the intention to report, meaning that the less negative they felt after the game, the more likely they were to report the troll – a finding contrary to our initial prediction in H2 that *stronger* negative emotions would lead to a *higher* intention to report. Although troll competence did not produce a significant effect, victim competence (1.86 [0.20, 3.52]) had a positive relationship with the intention to report, as predicted: the more competent the victim was perceived as being, the more likely the participant was to report the troll. Therefore, H3a – the prediction that increased perceived troll competence would lead to a lower intention to report – is rejected, but H3b – the prediction that increased perceived victim competence would lead to a higher intention to report – is supported. Finally, perceived group cohesion also had a negative effect on the intention to report a troll post-game (-2.77 [-4.69, -0.84]). The lower the perceived group cohesion score a participant gave the team, the more likely

they were to report the troll at game's end. Thus, H4 – that lower group cohesion will lead to increased intention to report – is supported.

Explanations for Reporting Choice Made. As mentioned earlier, when we asked participants whether they would have reported any of the players in game, we gave them the opportunity to explain the reasoning behind their decision. Only participants in the trolling condition ever said that they would report a player, but of those 31 participants who indicated that they would have reported a player, not all of them gave the same reason why. The vast majority included something to the effect of "Yes, the person that was flaming is making the game less fun for the others" (P8, male, age 25), mentioning either "offensive talk" (P19, female, age 20) or being "toxic" (P23, male, age 18). However, another theme that arose was the idea of mechanical skill at the game. Participant 64 (male, age 23) explains: *"*Knowing that I play with less experienced teammates I wouldnt [sic] report them. But if this was a real game in ranked I wouldve [sic.] reported both, for being unskilled.*"* This sentiment is echoed by Participant 118 (male, age 24), who explained that the confederates "felt very robotic and didn't have an idea of how to last hit or even show presence in the map." Last-hitting and map presence are examples of basic skills players have to execute in a game of League of Legends. Two other participants (P84, male, age 22; P116, male, age 20) also mentioned the difference between the experimental game and a ranked game, either saying directly or strongly implying that they would have reported the players if the game was ranked, as this would have affected their ability to win.

Taken together, these explanations would suggest that, although some participants do report based in a desire to bring justice by punishing those who flame, others do so for more selfish reasons: to increase their chance of winning. There are at least some players for whom ranked play and perceived competence are critical to the decision to report, but this is not the case for the entire community, and it was only a small part of our sample in the present study.

Discussion

The goal of the present study was to explore the antecedents of reporting behaviour in an online, in-game trolling situation. Our research question – how often do gamers report non-trolls (RQ1) – was answered by achieving complete separation in the dataset. None of our participants engaged in reporting in the non-trolling conditions. Our first expectation, however, was that the presence of an agent victim, operationalized in the present study as having sent in-game messages to the troll, would increase the likelihood that a bystander would report the troll when trolling was taking place (H1). This hypothesis was rejected based on our results. Whether the victim confederate spoke to the troll confederate had no discernible effect on the bystander's decision to report the troll. This could be an effect of the scripting of our trolling conditions. In order to ensure that victim chat remained consistent even when there was no trolling present, our victim and troll scripts were crafted to stand alone or be combined. As such, our victims spoke more neutrally than the average trolling victim (see Cook et al., 2019). Future studies varying the type of response a victim gives – trolling back, raging, seeking support – could determine whether our lack of an effect was due to the tameness of our victim's script, or whether victim agency simply does not impact reporting behaviour in online games.

Our second hypothesis (H2) predicted that trolling would lead to a more negative emotional state in-game, which would lead to a higher likelihood of the player's intention to report when trolling was occurring. Our results, however, do not support that conclusion, as although negative emotional state did predict the intention to report, it was not the expected way. Instead of more anger predicting a report, less negative emotion increased the chance that the participant would report the troll. This was a wholly unexpected result, and one that requires further investigation to completely explain. One possibility is that too much negative emotion

clouds the reason required to properly evaluate the situation and decide upon the best course of action. This is alluded to in Winterich and colleagues' (2015) study of moral emotions like disgust and happiness affecting ethical and unethical judgment. According to this study's results, if the bystander is disgusted enough, then they rely heavily on the perceived impact of the offense when making their judgment. Although trolling is decidedly unpleasant to experience, it is not typically life-threatening, and was not so in the present study. It could be that the stakes were simply not high enough for an excess of negative emotion to lead to reporting. However, as previously stated, this would require additional research to either confirm or deny.

Our third hypothesis took two parts: H3a) that trolling would lead to a low evaluation of the troll's competence, consequently increasing the likelihood of end-game reporting, and H3b) that when trolling occurs, a victim speaking up would lead to a higher evaluation of the victim's competence, consequently increasing the likelihood of end-game reporting. Although neither perception of competence was affected by victim agency, perceived victim competence did predict the intention to report on its own. Perceived group cohesion, too, affected the intention to report, but not as a function of victim agency like we predicted in H4. Together, these results would suggest that, on the whole, effective teamwork is more important to League of Legends players than individual performance, at least when it comes to reporting. In other words, there are two conditions that make a troll worthy of reporting: 1) if they are berating an effective teammate, or 2) if they are disrupting the performance of the team as a whole. This comes back to the principle of deservedness discussed by Hagai and Crosby (2016). Based on these results, one could infer that one only becomes deserving of punishment if the person's behaviour impacts the team's performance. However, troll warmth also achieved complete separation in our analyses; if the troll was perceived unkindly enough, they were automatically reported. Thus, it

would seem that group cohesion is as much a matter of morale as it is a matter of winning the game.

Theoretical Implications

A major theoretical implication of the present study is that the gaming context appears to be unique within the world of reporting offensive content, but there is still significant overlap with whistleblowing – the closest analogue to in-game reporting. Of the predictors tested in the present study, group cohesion – adapted from the concept of loyalty in whistleblowing literature (Misch et al., 2018; Waytz et al., 2013) – equity, and moral emotions, all of which have been shown to be predictive of whistleblowing (e.g., Henik, 2015), were equally predictive of reporting a troll. The moral emotion of anger, however, worked in the opposite direction: it was less negative emotion that led participants to intend to report, not more. Cook and colleagues (2018), much like Cheng and colleagues (2017), found that trolling is often driven by players being 'on tilt', referring to being in a negative state of mind during a game. This is often caused either by entering a losing streak or by having trolls present in an earlier game, something that was further supported by Cook and colleagues' (2019) results. In short, negative emotions are a key motivator that perpetuates the cycle of trolling (victims becoming perpetrators), but it has the opposing effect on the motivation to *stop* the trolling cycle (intervene or report). In the gaming context, it would appear that teamwork, effective or ineffective, is what ultimately decides whether a player will report another player. This marked difference between the gaming context and the business context (whistleblowing) serves as a call to expand our understanding of reporting behaviour, as only bits and pieces of both contexts seem generalizable, and the removal of offensive content is important all over the internet, not just in the gaming sphere.

The present study has also highlighted the complex way in which the principle of deservedness (Hagai & Crosby, 2016) applies in the online gaming context. Though League of Legends is a competitive, performance-driven context, it was the troll's warmth, not their competence, that achieved complete separation in our analyses. If the bystander thought the troll was unkind enough, they would be automatically reported, irrespective of their in-game performance or perceived competence. According to our results, the troll's behaviour was not enough to cause a report on its own (although there was never an intention to report anyone when no trolling was present). In order to report, our results would suggest that naïve participants had to conclude that the troll's actions were negatively affecting the cohesion of the team and were thus a risk to the team losing. The troll only fulfils the requirements of the principle of deservedness under two circumstances: if their unkindness crosses a certain threshold of acceptability, and if the entire team's performance or morale is negatively affected by their trolling. This corresponds at least in part to what whistleblowing literature has found (e.g., Andon et al., 2018), but we do not know if this trend extends onto other platforms such as Reddit (e.g., Paananen & Reichl, 2019) or YouTube (e.g., McCosker, 2014) where trolling behaviours are equally prevalent, but the competitive element is not there. Future studies should also explore the different thresholds of acceptability in terms of warmth or lack thereof in other communities, as it is possible that this varies depending on the platform in question. That said, the present study has given additional nuance to our understanding of how the principle of deservedness is applied when it comes to enacting justice in a virtual world.

Finally, our study has shown the potential of team cohesion and warmth in interventions designed to encourage reporting behaviour. As previously stated, a major portion of trolling literature is dedicated to prevention strategies like filtering and automatic troll detection (e.g.,

Cheng et al., 2017). However, as these methods are still imperfect, online communities need something to deal with trolling in the meantime. Cook and colleagues (2018) noted that, for the social norms of an online community to change, one of the key components required are what Heckathorn (1988) calls ‘enforcement resources’, which is to say a means to uphold the rules and regulations of the community. Reporting is, for now, the most effective enforcement resource League of Legends players have, but it is allegedly being underused (Cook et al., 2018). Based on this study, we now know that team morale (a combination of individual warmth and group cohesion) is an important factor in deciding whether or not to engage in justice behaviour (i.e., reporting). If industry and researchers alike want to encourage reporting and discourage trolling, interventions should be designed that emphasize team morale and camaraderie, as these appear to be the most important variables in the decision process.

Limitations & Future Directions

Although the present study is an important step forward in understanding what motivates reporting behaviour, particularly among bystanders, it is not without its limitations. First among these is the lack of behavioural data from the game itself. As mentioned earlier, due to an update change, we were unable to retrieve our game data, leaving us with only self-report measures instead of our intended combination of objective and subjective measures. In addition, there is the question of ecological validity that comes with a scripted trolling situation. Due to ethical limitations, we were unable to use flaming statements that match the degree of severity and intensity that are common in League of Legends (see Cook et al., 2019 for examples). Confederates were also unable to react naturally to participants’ messages, as their priority had to be preserving the sameness of trolling conditions across participants. There was also no explicit manipulation check for victim agency (if the victim talked or not), and so we cannot be

completely sure that participants noticed the victim's chat or not. However, participants' explanations for reporting frequently mentioned the victim's silence or otherwise, suggesting that it was noticed at least by some. Finally, due to the length of the data collection phase and the champion rotation system inherent to League of Legends, we were not able to have the confederates play the same characters for all participants. Although all possible steps were taken to reduce the variation (e.g., confederates playing the same *type* of champion, even if it could not be the exact same one), it would have been ideal to use computerized bots that played an identical game each time. Future researchers could potentially reach out to game companies and collaborate directly with the company to make this a possibility.

Despite these limitations, the present study still serves as a foundation for future research on reporting norms in online games and other locales in cyberspace. Cross-platform research as a whole is sorely lacking in the field of trolling (see Cook et al., 2018), and based on the present study's results, there is a high likelihood that factors that differ between platforms, such as the degree of competition involved, could cause differences in reporting norms. Without the competitive element, it is possible that a simple warmth threshold may be enough to predict reporting behaviours, based on our current results. A major opportunity for future research is thus to compare between reporting norms on different platforms. The present study confirms that not all reporting literature can be generalized across contexts, but we still know little about the specific mechanisms behind these differences, e.g., is it the anonymity or the competitiveness of online games that makes reporting norms different from those in the business context of whistleblowing? Future studies could also perform more qualitative, interview-based studies to gain deeper insight into the specific motivations behind reporting beyond competition and a personal desire to win.

In conclusion, there is still much to do in order to completely understand what motivates bystanders to report trolls. The present study has opened the door to more non-troll-centric studies of trolling interactions and given new perspective on bystanders' role in the trolling cycle. With additional studies, there is the potential to determine the generalizability of the present results and expand the trolling literature on intervention, not only simple detection and filtering. By performing cross-platform studies and further examining the variables involved in reporting trolls to authorities, researchers can increase our understanding of this online phenomenon, and perhaps learn what we need to prevent the instances in which it becomes damaging to our online communities.

References

Abbott, N., & Cameron, L. (2014). What makes a young assertive bystander? The effect of intergroup contact, empathy, cultural openness, and in-group bias on assertive bystander intervention intentions. *Journal of Social Issues*, *70*, 167-182. doi: 10.1111/josi.12053

Allison, K.R., & Bussey, K. (2016). Cyber-bystanding in context: A review of the literature on witnesses' responses to cyberbullying. *Children and Youth Services Review*, *65*, 183-194. doi: 10.1016/j.childyouth.2016.03.026

Andon, P., Free, C., Jidin, R., Monroe, G.S., & Turner, M.J. (2018). The impact of financial incentives and perceptions of seriousness on whistleblowing intention. *Journal of Business Ethics*, *151*, 165-178. doi: 10.1007/s10551-016-3215-6

Alman, J., & McKay, D. (2017). Theoretical foundations of team matchmaking. *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent systems (AAMAS 2017), May 8-12, Sao Paulo, Brazil*, 1073-1081. Retrieved from http://www.ifaamas.org/Proceedings/aamas2017/pdfs/p1073.pdf

Balci, K., & Salah, A.A. (2015). Automatic analysis and identification of verbal aggressive and abusive behaviors for online social games. *Computers in Human Behavior, 53*, 517-526. DOI: 10.1016/j.chb.2014.10.025

Bastiaensens, S., Vandebosch, H., Poels, K., van Cleemput, K., DeSmet, A., & de Bourdeaudhuij, I. (2014). Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. *Computers in Human Behavior, 31*, 259-271. doi: 10.1016/j.chb.2013.10.036

Baumert, A., & Schmitt, M. (2016). Justice sensitivity. In C. Sabbagh & M. Schmitt (Eds.) *Handbook of Social Justice Theory and Research* (pp. 161-180). New York, NY: Springer.

Bennett, S., & Banyard, V.L. (2016). Do friends really help friends? The effect of relational factors and perceived severity on bystander perception of sexual violence. *Psychology of Violence, 6*(1), 64-72. doi: 10.1037/a0037708

Blackburn, J., & Kwak, H. (2014, April). *STFU NOOB! Predicting crowdsourced decisions on toxic behavior in online games*. Paper presented at the 23rd International World Wide Web conference, Seoul, South Korea.

Blizzard. (2004). World of Warcraft [Computer software]. Retrieved from https://us.shop.battle.net/en-us/product/world-of-warcraft-subscription?utm_source=Google&utm_medium=Search&utm_content=29721375&utm_campaign=BLZ_WoW_EG_Modern-Classic_UA_2020_NA

van den Bos, K., & Bal, M. (2016). Social-cognitive and motivational processes underlying the justice motive. In C. Sabbagh & M. Schmitt (Eds.), *Handbook of Social Justice Theory and Research* (pp. 181-200). New York, NY: Springer.

Brody, N., & Vangelisti, A.L. (2016). Bystander intervention in cyberbullying. *Communication Monographs, 83*(1), 94-119. doi: 10.1080/03637751.2015.1044256

Burkholder, R. (2019, August 7). *Co-constructing virtual identities: Insights from linguistic analysis*. Paper presented at the annual meeting of the Digital Games Research Association (DiGRA), Kyoto, Japan.

Carroll, J.M. … Van Metre, C. (2001). Designing our town: MOOsburg. *International Journal of Human-Computer Studies, 54*, 725-751. doi: 10.1006/ijhc.2000.0438

Cassematis, P.G., & Wortley, R. (2013). Prediction of whistleblowing or non-reporting observation: The role of personal and situational factors. *Journal of Business Ethics, 117*, 615-634. doi: 10.1007/s10551-012-1548-3

Chen, C.-P., & Lai, C.-T. (2014). To blow or not to blow the whistle: The effects of potential harm, social pressure and organisational commitment on whistleblowing intention and behaviour. *Business Ethics: A European Review*, *23*, 327-342. doi: 10.1111/beer.12053

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizili, C., & Leskovec, J. (2017). *Anyone can become a troll: Causes of trolling behavior in online discussions*. Paper presented at the 20th ACM Conference on Computer-Supported Cooperative Work and Social Computing, Portland, Oregon.

Chesney, T., Coyne, I., Logan, B., & Madden, N. (2009). Griefing in virtual worlds: Causes, casualties and coping strategies. *Information Systems Journal*, *19*, 525-548. doi: 10.1111/j.1365-2575.2009.00330.x

Colp, T. (2020, March 31). When is BlizzCon 2020? It may be sooner than you think! *Blizzard Watch*. Retrieved from https://blizzardwatch.com/2020/03/31/blizzcon-2020-date/

Cook, C., Schaafsma, J., & Antheunis, M.L. (2018). Under the bridge: An in-depth examination of trolling in the online gaming context. *New Media & Society*, *20*, 3323-3340. doi: 10.1177/1461444817748578

Cook, C., Conijn, R., Antheunis, M.L., & Schaafsma, J. (2019). For whom the gamer trolls: A study of trolling interactions in the online gaming context. *Journal of Computer-Mediated Communication*. Advance online publication. doi: 10.1093/jcmc/zmz014

Darley, J.M., & Latané, B. (1968). Bystander intervention in emergencies: Diffusion of responsibility. *Journal of Personality and Social Psychology*, *8*, 377-383. doi: 10.1037/h0025589

Dasborough, M.T., Hannah, S.T., & Zhu, W. (2019). The generation and function of moral emotions in teams: An integrative review. *Journal of Applied Psychology*. Advance online publication. doi: 10.1037/apl0000443

Derek. (2013, March 30). *Improving player behavior in League of Legends* [Video]. YouTube. https://www.youtube.com/watch?v=a-5yKdpR0kU

Dibbell, J. (1993, 23 December). A rape in cyberspace, or how an evil clown, a Haitian trickster spirit, two wizards, and a cast of dozens turned a database into a society. *The Village Voice*. Retrieved from http://www.juliandibbell.com/texts/bungle_vv.html

Dourish, P. (1998). Introduction: The state of play. *Computer Supported Cooperative Work: The Journal of Collaborative Computing, 7,* 1-7. Retrieved from https://link.springer.com/article/10.1023%2FA%3A1008697019985?LI=true

Ellemers, N., van der Toorn, J., Paunov, Y., & van Leeuwen, T. (2019). The psychology of morality: A review and analysis of empirical studies published from 1940 through 2017. *Personality and Social Psychology Review, 23*, 332-366. doi: 10.1177/1088868318811759

Erickson, T. (1993). *From interface to interplace: The spatial environment as a medium for interaction*. Paper presented at the European Conference on Spatial Information Theory, Marciana Marina, Elba, Italy. Retrieved from https://link.springer.com/chapter/10.1007%2F3-540-57207-4_26?LI=true

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. London, UK: Sage.

Fischer, P., Krueger, J.I., Greitemeyer, T., Vogrincic, C., Kastenmüller, A., Frey, D., … & Kainbacher, M. (2011). The bystander-effect: A meta-analytic review on bystander

intervention in dangerous and non-dangerous emergencies. *Psychological Bulletin*, *137*, 517-537. doi: 10.1037/a0023304

Fiske, S.T., Cuddy, A.J.C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878-902. doi: 10.1037//0022-3514.82.6.878

Fox, J., Gilbert, M., & Tang, W.Y. (2018). Player experiences in a massively multiplayer online game: A diary study of performance, motivation, and social interaction. *New Media & Society*, *20*, 4056-4073. doi: 10.1177/1461444818767102

Fox, J., & Weisberg, S. (2019). *An {R} companion to applied regression* (3rd ed). Thousand Oaks, CA: Sage.

Gallegos, A. (2012, October 16). Riot Games releases awesome League of Legends infographic. *IGN*. Retrieved from https://www.ign.com/articles/2012/10/15/riot-games-releases-awesome-league-of-legends-infographic

Gollwitzer, M., & van Prooijen, J.-W. (2016). Psychology of Justice. In C. Sabbagh & M. Schmitt (Eds.), *Handbook of Social Justice Theory and Research* (pp. 61-82). New York, NY: Springer.

Hagai, E.B., & Crosby, F.J. (2016). Between relative deprivation and entitlement: An historical analysis of the battle for same-sex marriage in the United States. In C. Sabbagh & M. Schmitt (Eds.) *Handbook of Social Justice Theory and Research* (pp. 477-489). New York, NY: Springer.

Harari, G.M., Graham, L.T., & Gosling, S.D. (2015). Personality impressions of World of Warcraft players based on their avatars and usernames: Consensus but no accuracy.

International Journal of Gaming and Computer-Mediated Simulations, 7, 58-73. doi: 10.4018/IJGCMS.2015010104

Hayes, A.F. (2017). *Introduction to mediation, moderation, and condition process analysis* (2nd ed.). Guilford press.

Heckathorn, D.D. (1988). Collective sanctions and the creation of prisoner's dilemma norms. *Journal of Sociology, 94,* 535-562. Retrieved from http://www.jstor.org/stable/2780253

Henik, E. (2015). Understanding whistle-blowing: A set-theoretic approach. *Journal of Business Research, 68,* 442-450. doi: 10.1016/j.jbusres.2014.06.004

Herring, S., Job-Sluder, K., Scheckler, R., & Barab, S. (2002). Searching for safety online: Managing "trolling" in a feminist forum. *The Information Society, 18,* 371-384. doi: 10.1080/01972240290108186

Hudson, B. (2014). Funny Games: Understanding videogames as slapstick and the experience of game-worlds as shared cultural references. In D. Stobbart and M. Evans (Eds.), *Engaging with Videogames: Play, Theory and Practice* [E-reader version] (pp. 109-120). DOI: https://doi.org/10.1163/9781848882959_011

Jacobson, D. (2001). Presence revisited: Imagination, competence, and activity in text-based virtual worlds. *CyberPsychology & Behavior*, *4*(6), 653-673. doi: 10.1089/109493101753376605

Kahn, A.S., & Williams, D. (2016). We're all in this (game) together : Transactive memory systems, social presence, and team structure in multiplayer online battle arenas. *Communication Research, 43,* 487-517. DOI: 10.1177/0093650215617504

Kwak, H., Blackburn, J., & Han, S. (2015, April). *Exploring cyberbullying and other toxic behavior in team competition online games*. Paper presented at the CHI : Crossings conference, Seoul, South Korea.

Latan, H., Jabbour, C.J.C., & de Sousa Jabbour, A.B.L. (2019). To blow or not to blow the whistle: The role of rationalization in the perceived seriousness of threats and wrongdoing. *Journal of Business Ethics*. Advance online publication. doi: 10.1007/s10551-019-04287-5

League of Legends Basics. (2019). *Game Modes*. Retrieved from https://lolinform.weebly.com/game-modes.html

Lee, S.J., Jeong, E.J., & Jeon, J.H. (2019). Disruptive behaviors in online games: Effects of moral positioning, competitive motivation, and aggression in “League of Legends”. *Social Behavior and Personality: An International Journal, 47*, 1-9. doi: 10.2224/sbp.7570

MacKenzie, K. R. (1983). The clinical application of a group climate measure. In R. R. Dies & K. R. MacKenzie (Eds.), *Advances in group psychotherapy: Integrating research and practice* (pp. 159-170). New York: International Universities Press.

McCosker, A. (2014). Trolling as provocation: YouTube’s agonistic publics. *Convergence: The International Journal of Research into New Media Technologies, 20*, 201-217. doi: 10.1177/1354856513501413

McGlynn, J., & Richardson, B.K. (2014). Private support, public alienation: Whistle-blowers and the paradox of social support. *Western Journal of Communication, 78*, 213-237. doi: 10.1080/10570314.2013.807436

Misch, A., Over, H., & Carpenter, M. (2018). The whistleblower's dilemma in young children: When loyalty trumps other moral concerns. *Frontiers in Psychology, 9,* 1-9. doi: 10.3389/fpsyg.2018.00250

Mock, B. (2019, August). What new research says about race and police shootings. *CityLab*. Retrieved from https://www.citylab.com/equity/2019/08/police-officer-shootings-gun-violence-racial-bias-crime-data/595528/

Mora-Cantallops, M., & Sicilia, M.-A. (2018a). Exploring player experience in ranked League of Legends. *Behaviour & Information Technology, 37,* 1224-1236. DOI: 10.1080/0144929X.2018.1492631

Mora-Cantallops, M., & Sicilia, M.-A. (2018b). MOBA games: A literature review. *Entertainment Computing, 26,* 128-138. doi: 10.1016/j.entcom.2018.02.005

Mora-Cantallops, M., & Sicilia, M.-A. (2019). Team efficiency and network structure: The case of professional League of Legends. *Social Networks, 58,* 105-115. doi: 10.1016/j.socnet.2019.03.004

Montada, L., & Maes, J. (2016). Justice and self-interest. In C. Sabbagh & M. Schmitt (Eds.), *Handbook of Social Justice Theory and Research* (pp. 109-126). New York, NY: Springer.

Munro, I. (2017). Whistle-blowing and the politics of truth: Mobilizing 'truth games' in the WikiLeaks case. *Human Relations, 70,* 519-543. doi: 10.1177/0018726716672721

Paananen, A., & Reichl, A.J. (2019). Gendertrolls just want to have fun, too. *Personality and Individual Differences, 141,* 152-156. doi: 10.1016/j.paid.2019.01.011

Penev, B. (2020, April 3). European Masters 2020 Spring begins Monday. *EarlyGame*. Retrieved from https://www.earlygame.com/european-masters-2020-spring-begin-monday/

Postmes, T., Spears, R., & Lea, M. (1998). Breaching or building social boundaries? SIDE-effects of computer-mediated communication. *Communication Research, 25,* 689-715. DOI: 10.1177/009365098025006006

R Core Team (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.R-project.org/.

Riot Games (2009). League of legends [Computer software]. Retrieved from http://euw.leagueoflegends.com/

Riot Games (2020). Terms of service. *Riot Games*. Retrieved from https://www.riotgames.com/en/terms-of-service

RzK. (2013). Any good gaming subreddits? Retrieved from https://www.reddit.com/r/AskReddit/comments/19t0t5/any_good_gaming_subreddits/

Slonje, R., Smith, P.K., & Frisén, A. (2012). The nature of cyberbullying, and strategies for prevention. *Computers in Human Behavior*, *29*, 26-32. doi: 10.1016/j.chb.2012.05.024

Thacker, S., & Griffiths, M.D. (2012). An exploratory study of trolling in online video gaming. *International Journal of Cyber Behavior, Psychology and Learning*, *2*, 17-33. doi: 10.4018/ijcbpl.2012100102

Valentine, S., & Godkin, L. (2019). Moral intensity, ethical decision making, and whistleblowing intention. *Journal of Business Research*, *98*, 277-288. doi: 10.1016/j.jbusres.2019.01.009

Venables, W.N., & Ripley, B.D. (2002). *Modern Applied Statistics with R* (4th ed.). New York, NY: Springer.

Verheij, T., Bleize, D., & Cook, C. (*in press*). Friendly fire off: Does cooperative gaming in a competitive setting lead to prosocial behaviours? *Press Start.*

Watson, D., Clark, L.A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scale. *Journal of Personality and Social Psychology*, *54*(6), 1063-1070. doi: 10.1037/0022-3514.54.6.1063

Waytz, A., Dungan, J., & Young, L. (2013). The whistleblower's dilemma and the fairness-loyalty tradeoff. *Journal of Experimental Social Psychology*, *49*, 1027-1033. doi: 10.1016/j.jesp.2013.07.002

Wilcox, R.R. (2005). *Introduction to robust estimation and hypothesis testing* (2nd ed.). Burlington, MA: Elsevier.

Winterich, K.P., Morales, A.C., & Mittal, V. (2015). Disgusted or happy, it is not so bad: Emotional mini-max in unethical judgments. *Journal of Business Ethics*, *130*, 343-360. doi: 10.1007/s10551-014-2228-2

Chapter 6: Conclusion

Conclusions and Theoretical Implications

The global aim of the present dissertation is to determine by exploring the perspectives of all actors in trolling interactions (trolls, victims, and bystanders) RQ1) what constitutes trolling, RQ2) what motivates trolls, and RQ3) how community response (victims and bystanders) react to different types of trolling in differing contexts. In Chapter 2, I uncovered the perspective of the troll by interviewing self-confessed trolls, something only rarely done at the time of publishing (see Phillips 2011; 2013). This study also allowed me to examine the motivations of trolls directly without having to resort to interpreting behaviour to deduce what the motivation behind the behaviour might be. We learned that trolling can consist of both verbal and behavioural antics, and that these are designed to provoke victims, either into conversation or into rage, depending on the motivation (RQ1). The motivations behind trolling can also vary, consisting primarily of personal enjoyment, boredom reduction, being 'tilted' (in a negative headspace), or even interaction-seeking (RQ2). In essence, trolls can use their words, game mechanics, and their own inventiveness to have fun, to make friends, to conduct their own personal social experiments, or to vent their rage, all at another person's expense. My results in Chapter 2 would suggest that trolling consists of multiple instrumental behaviours designed to achieve a pleasurable end for the troll, and that these can be largely grouped into verbal and behavioural trolling (RQ1). These results would also suggest that the motivation behind the trolling can vary in antisocial- or prosocial-ness, but that it is never simple wonton destruction, as posited by extant literature (e.g., Buckels, Trapnell, & Paulhus, 2014).

The next three chapters, although providing some subtle nuances to my answers to RQ1 and RQ2, focused mainly on answering RQ3: the impact of the community. I took a broader

perspective in Chapter 3 when I looked at the chat logs from over 10,000 reported cases of online trolling. This chapter took the question of human perspective out of the equation and examined full trolling interactions through the lens of machine learning to see what kind of chat is considered trolling, and how victims and bystanders respond to trolling in the wilds of cyberspace. What we found is that bystanders and victims often participate in the trolling, escalating the situation beyond the troll's initial provocation. In addition, we found that they often begin trolling themselves, engaging in negatively-valenced emotionally-charged banter and profanity that is semantically difficult if not impossible to distinguish from the initial troll's offense. Chapter 4 was focused on bringing the victim's perspective on trolling to the forefront, while also exploring how one's offline cultural context could potentially impact the experience of being a trolling victim. This study added extra nuance, showing that the cultural context of a person and the type of trolling experienced can interact to create the myriad victim reactions we can find in extant literature. Contrary to what we expected, our Taiwanese participants were the most aggressive and retaliatory, while our Pakistani participants tended try to build and repair relationships with our pre-programmed confederates. This explains at least partially why victims and bystanders are so inconsistent in their responses in trolling literature, leading to the variety of trolling interactions showcased in earlier research (e.g., Herring, Job-Sluder, Scheckler, & Barab, 2002; Thacker & Griffiths, 2012). Finally, in Chapter 5, I examined the bystander's perspective on trolling and took a critical look at how a game's environment and culture – League of Legends, in this case – could impact bystanders' response to witnessing another player being trolled. The results of this study revealed that although bystanders have the power to change the trolling interaction, they seldom use it for trolling alone. In order for a report to be made, the trolling had to have negatively affected the perceived group cohesion, with rates being boosted

by a victim who is perceived as being competent. In other words, unless the trolling affected the entire team atmosphere and teamwork, trolling by itself was not enough to merit a report, at least amongst our participants. Thus, to answer RQ3, the community's possible responses are as varied as trolls' possible trolling behaviours, and they can both promote or inhibit future trolling, depending on the victims' and bystanders' cultural background and the game they are playing.

However, throughout these studies, I not only fulfilled the goal of the dissertation, but also came to several important conclusions that have significant implications for the field of trolling research. These implications can also affect existing theories of online behaviour and aggression, over and above what we have learned about trolling.

Defining Trolling Across Communities. While gathering all of the different perspectives of the various actors in trolling interactions, it became clear that what constitutes trolling can differ dramatically depending on who you ask. There are a few standbys that show up repeatedly – flaming, feeding (purposely getting oneself killed in-game), and using racist or sexist language, to name a few – but outside of these, there is a considerable variety of possible trolling behaviours that arise when you talk to trolls (Chapter 2) and when look at trolling interactions (Chapter 3). While verbal trolling is particularly prominent, this can consist of almost everything from locker-room-style trash-talking to screeching into a microphone repeatedly. Behavioural trolling is even more varied, as it depends on the mechanics available to the troll; this means that the types of behavioural trolling are limited only by the types of multiplayer games that exist. However, it is not only the platforms that can cause variation: the perspective taken also add considerable variation. This is because what one person calls trolling, another may consider completely innocuous, much like flaming (see O'Sullivan & Flanagin, 2003). In other words, if the troll considers their actions to be successful trolling, that does not

guarantee that the victim will feel trolled, or that bystanders would report the behaviour as trolling.

This has important implications for the question of trolling's definition, an ongoing debate in the field (see Chapters 1 and 2). Throughout the course of my studies, I developed the following definition of trolling: trolling is the instrumental exploitation of game, chat, or website mechanics, at another person's expense. However, those familiar with trolling research will likely notice that this is a much broader definition than those offered in most other scientific articles on the subject (e.g., Buckels, Trapnell, & Paulhus, 2014; Fichman & Sanfilippo, 2014; Thacker & Griffiths, 2012). This is intentional, as I have come to the conclusion that due to the inherent question of perspective, trolling can never have a true, single definition (see Sanfilippo, Yang, & Fichman, 2017). What I bring in this dissertation is a base definition that can be modified to suit the individual communities in which trolling occurs. There is no one-size-fits-all list of trolling behaviour that can determine 100% of the time what is or is not trolling; the only constants I found in my research, both in my own studies and the extant literature, are that a) trolls act the way they do for a reason, with a goal in mind, b) the main limitations that trolls face are those of the programming they work with, and c) all trolling includes a perceived victim, whether it is the troll, a bystander, or the victim themselves that perceives one. Thus, the definition that I present here includes instrumentality, mechanics, and a victim as its only components; the specifics should change depending on the community both on- and offline (see Chapter 4) in question, and who in that community is talking about the trolling occurring (see Chapters 2 and 5). That said, I do not believe that this prevents trolling researchers from working together as a unit to grow our understanding of trolling as a phenomenon; it simply means that not everything we find will generalize across platforms.

The Trolling Cycle. Another finding that has significant implications for future trolling research is the identification of a possible trolling cycle. This is something that was already established in literature concerning trolling's sister behaviour – cyberbullying (Chapin & Coleman, 2017) – but had only ever been hinted at in existing trolling literature (see Bishop, 2014; Thacker & Griffiths, 2012). What this means is that, in all likelihood, people who are trolls today started out as victims of other trolls, just as victims of cyberbullying are more likely to become cyberbullies in the future (Chapin & Coleman, 2017). The most powerful evidence for this is in Chapter 2, where the entire sample of trolls reported being victims before their first instance of engaging in trolling behaviour. However, we can also see evidence for a trolling cycle in Chapter 3. In that chapter, I found that the chat of trolls and their in-game teammates – those most-exposed to their trolling – share many characteristics traditionally associated with trolls, such as a general negative emotional valence and the use of profanity, suggesting that victims and bystanders can start to troll within a single game's worth of exposure to another person trolling. We also saw rage and troll-like actions from our participants in Chapter 4, and they were being trolled by a script as opposed to a real human player.

If the trolling cycle really does progress so much quicker than the cyberbullying cycle, which takes many repeated exposures to cyberbullying to take effect (see Chapin & Coleman, 2017), this opens up new avenues of research for scholars. Though the studies in this present dissertation suggest the existence of the cycle, they do not specify the mechanisms behind the cycle. All of my studies take place in an online gaming context, so it remains to be seen how the cycle may progress on different platforms. Most cyberbullying studies focus on social media (e.g., Bastiaensens et al., 2014; Vandebosch & van Cleemput, 2009), so if the platform impacts cycle speed, this may be why there appears to be such a stark speed difference between the two

cycles. Another possible mechanism behind the trolling cycle could be anonymity. Although both trolling and cyberbullying both take place online, while cyberbullies tend to be known to their victims offline (e.g., Vandebosch & van Cleemput, 2009), the opposite is usually true of trolls in online games (see Chapter 2). This means that, in the majority of cases, victims do not have to fear offline retaliation if they fight back, which could also account for the difference in cycle speeds. Overall, my results would suggest that if researchers are primarily interested in victims of trolling, they should target their manipulations at the earliest point possible in the trolling interaction, as the victim may have changed into a troll before the conversation's end.

This finding of a trolling cycle also opens up questions about online aggression more generally. In Chapter 3, we found that having a troll on one's team seems to 'corrupt' the entire chat process, making trolls and victims almost indistinguishable from one another in terms of sentiment analysis, and this despite sentiment being a key variable used in previous research to distinguish trolls from others in an interaction (e.g., Cheng, Bernstein, Danescu-Niculescu-Mizil, & Leskovec, 2017). However, the SIDE model's anonymity is only one possible explanation for this effect. Tedeschi, Smith, and Brown's (1974) theory of coercive action, for instance, would suggest that trolling is a form of noxious stimulus used to manipulate others into giving the troll what they want from an interaction. This is in line with the more selfish trolling motivations captured in Chapter 2. However, this trolling cycle could also be reflecting the adaptation process of a person entering into a new culture with a new set of norms. When trolls act and talk the way they do, whether it is intentional or not, they are also serving as models of what kind of behaviour is acceptable in the gaming context (Bandura, 1978). It could be that the trolling cycle is particularly rapid in games especially because of the competitive nature of the online gaming studied in this dissertation; users are particularly pressured to adapt to the local norms, whether it

be via a coercion tactic (Tedeschi et al., 1978) or the simple pressure to win (Mora-Cantallops & Sicilia, 2018). This adaptation to local norms and social learning perspective would also explain why the Taiwanese participants acted so apparently contrary to existing theories of cross-cultural psychology and aggression (Leung & Cohen, 2011; Markus & Kitayama, 1991) in Chapter 4, retaliating when most existing literature on face-valuing cultures would predict avoidance. They were perhaps not acting out of character, but rather doing exactly what Hashimoto and Yamagishi (2013) would predict: adapting to the local norms to preserve face. In short, there is still a long way to go before we can understand the exact mechanisms behind the trolling cycle and how these gears and cogs may interact; for now, all we can know for certain is that the cycle exists.

The Complex Agents of Trolling Interactions. Finally, my dissertation highlights the complexity of each of the actors in trolling interactions, including the troll, who is often presented as a one-dimensional antagonist, especially in media (e.g., Stein, 2016). In Chapter 2, we found that trolls are motivated by a variety of things, not simply wonton destruction or the sheer desire to make others unhappy (e.g., Buckels et al., 2014). Trolls can troll to inspire conversation or interaction with others, or to alleviate boredom; in some cases, the only reason they troll is to give existing trolls the proverbial taste of their own medicine. Chapter 4 revealed the complexities of being a victim of trolling: that victims bring their own offline culture to the interaction, and can react in many different ways, everything from the rage trolls seem to desire (see Chapter 2) to seeking help and support from bystanders. Even bystanders themselves can make different choices based on individual differences and environment that can change trolling interactions, as hinted at in Chapter 3 and explored more fully in Chapter 5. Throughout this dissertation, my results have demonstrated that trolling is more than just one angry person

ruining someone's day, just as trolls are more than storybook villains. Trolling is a complex interaction between fully agent actors, a fact that should not be neglected in future research of this phenomenon.

My findings in terms of online and offline context also have interesting implications for trolling that involve other theories in media psychology. For instance, most of the chapters in this dissertation touch in some way on the social identity model of deindividuation effects (SIDE; Postmes, Spears, & Lea, 1998). In Chapter 3 in particular, the chat logs revealed that the amplification of emotion and polarization of thoughts and behaviour on a team with a troll that the SIDE model would predict. This gives us two pieces of information: 1) anonymity is an important factor to consider when researching trolling in online games, and 2) trolling in non-anonymous contexts, if it exists at all, is likely to be extremely different from what we have seen in the past four studies. Levels of retaliation should theoretically decrease, and it is possible that more culturally-normative responses would take over (e.g., Leung & Cohen, 2011; Markus & Kitayama, 1991). However, this is not the only psychological theory that can be applied to trolling successfully. In Chapter 2, we found that the motivations to troll were relatively consistent, despite our multicultural subject pool. The two primary triggers for trolling behaviour were universally 'tilt' (frustration) and boredom. This falls squarely in line with flow theory (Csikszentmihalyi, Abuhamdeh, & Nakamura, 2005), particularly when combined with the demand framework as presented by Bowman (2018). According to flow theory, when a task is too easy, a person becomes bored, while if a task is too difficult, they become frustrated (Csikszentmihalyi et al., 2005). In online games, there are many different kinds of tasks that must be completed: some cognitive (e.g., solving a puzzle in-game), some physical (e.g., reacting quickly to an enemy attack), some emotional (e.g., making a moral choice to advance a

storyline), and still others social (e.g., working in a team to accomplish an in-game goal). These are called demands (Bowman, 2018; Cook, 2019). In my 2019 article, I proposed that if any of these demands is insufficient or too much, boredom or frustration will result, and consequently, trolling. However, this could also be extended to victims and bystanders if one envisions trolling as the social and emotional demand that pushes victims and bystanders to react and try to reduce the demand. How exactly they react, however, is likely to depend on other factors, like their cultural context.

Depending on a person's conception of reputation (Leung & Cohen, 2011) – whether they adopt an honour code, try to save face, or defend their dignity – they react differently when they are trolled. In Chapter 4, we found that while our participants from face-valuing cultures were unusually aggressive online, our participants from honour-valuing cultures were much more avoidant. Our Dutch participants, representing dignity-valuing cultures, were the least reactive of all, acting nonchalant in the face of online trolling. This reaction also depended on the type of trolling, with flaming producing either aggression or avoidant, withdrawing behaviour depending on the cultural context of the victim, and ostracism producing primarily avoidance, irrespective of victim culture. Again, this gives us two key pieces of information: 1) trolling interventions designed to stop reciprocal aggression (trolling back) need to take offline cultural context into account, and 2) those same interventions will likely need to be adjusted depending on the specific kind of trolling the designer wants to prevent. From a theoretical perspective, our studies' results challenge the status quo of culture and aggression. It appears that the online context interferes with culturally-typical responses according to theories of cultural logic (Leung & Cohen, 2011) and self-construal (Markus & Kitayama, 1991). I already mentioned earlier that this could be due to the individual cultures interacting with the culture of the internet. It could,

however, simply be the effect of anonymity as postulated by the SIDE model (Postmes et al., 1998) that creates the results we find; maybe face concerns are mitigated by the fact that they need not fear retaliation in an anonymous setting, and maybe honour does not require defending when the threat is immaterial, in an online world where no one else will see it. Overall, my findings show the intricacies of trolling as a phenomenon – that trolling and the actors involved in it are complex and nuanced, and they need to be considered in research alongside the troll as a person.

Practical Implications

Because this dissertation focuses on trolling in online games, there is inherent interest in my results for the businesses behind the gaming industry. One such finding is the importance of the individual actors' perspectives to the definition of trolling and how the interaction plays out. Researchers in both user experience and academia alike need to take serious consideration into their perspective when they research trolling in order to capture the full picture of trolling, or answer specific questions related to trolling. My findings would suggest that what works to suppress trolling in one cultural context may not necessarily work in another. This cultural context can be referring to several things: one's offline culture, as in Chapter 4, or the culture of a game or platform, as in Chapter 5. Translated to the gaming industry, this means that an algorithm or intervention that deals with trolling effectively on one regional server may not work as well on a different regional server. This also means that what works on one game or genre of game to reduce trolling is not necessarily generalizable to all games. Because each game has its own existing culture, based on my research, each game is likely to have its own unique spin on the basic definition of trolling this dissertation provides. This means that both human and automatic (algorithm-based) content moderators will have to be specific to not only certain

games or genres, but also to particular regional servers in order to be truly effective at reducing the toxic kind of trolling.

Perhaps more important, however, is the finding in Chapter 2 that trolls believe reporting to be an effective trolling deterrent. This gives industry concrete feedback that their methods are working, at least in part. The trolls that get reported are confirming that this makes them less likely to troll again, but there are still plenty of trolls operating in games despite this fact. One idea that was presented by the trolls when asked how they would reduce trolling is to increase the amount of feedback given when reports are dealt with. They suggested that, when an offender was punished, the person who reported the offense should be notified, so that they know that their reporting behaviour was effective. We can see that this has already been implemented in some online games such as League of Legends, but it is not the only option, nor has this method been tested independently for effectiveness.

The results of Chapter 3 would suggest that the trolling cycle is a process that occurs extremely rapidly, transforming victims into trolls over the course of a single game. If this is the case, then having a report function at the end of the game, as is the case in most e-sports titles today (e.g., Overwatch, League of Legends, Dota 2), is probably too late to deal with trolling efficiently. If companies wish to stop the spread of trolling, they need to have some kind of functionality that deals with offenders, or at least gives bystanders and victims the perception that they are being dealt with, long before the game is over. Knowledge of the trolling cycle and the strengths and weaknesses of the reporting function is critical knowledge for companies who wish to expand their player base and provide a safe and playing environment for their customers. Though it is likely to weed out frequent but toxic players at the outset, in the long run, it is likely to be both the most effective and profitable strategy.

Future Directions

In addition to looking at new platforms outside of the online gaming realm, future studies should also be careful to take into account the duality of trolling: the "good" trolling (Paul, Bowman, & Banks, 2015) versus the trolling that is often referred to in research as "toxicity" (Kwak & Blackburn, 2014). Only in my first study, presented in Chapter 2, do I address the more positive kind of trolling. This consists of joking around with friends in a game or posting a largely inoffensive meme in a forum to make people laugh (see Paul et al., 2015). In Chapter 2, we see that not all trolls are malicious, or that that is not their primary intention, but this is the only time that I talk about the kind of trolling whose goal is fundamentally prosocial: to increase collective enjoyment of the game or social media platform. Researching this side of trolling may not provide the same degree of urgency as researching the more toxic side of trolling, as there is less risk involved for the general population in this more lighthearted trolling. However, it is still a necessary component to research in order to fully understand trolling as an internet phenomenon. Questions like "is humorous trolling contagious" and "what makes a trolling post funny versus offensive" still need to be answered.

In terms of opportunities for new research in the field of trolling, the present dissertation has shown the importance of social norms – of games (Chapters 3 and 5), of culture (Chapter 4), and even of trolls (Chapter 2) – in perpetuating or inhibiting trolling behaviour; the next logical step for researchers wanting to minimize trolling in online communities is to try and change these problematic norms. In Chapter 2, trolls were clear in saying that trolling had become normalized, particularly in the online gaming community. However, we still do not know how that process occurred; we do not yet know when flaming and shouting in all caps became acceptable online, despite its total unacceptability offline. Throughout the dissertation, we can

see that this is likely to do with anonymity and the polarization that happens in anonymous groups (Postmes et al., 1998). Future studies could test this theory more directly by altering the degree of anonymity on controlled social media or game platforms over time and seeing if and how the social norms change. Another option could be to develop training or interventions designed to combat the effects of anonymity, for example by teaching young children – the source of the next generation of online social norms – that the online world and its consequences are as real as the offline world. All of these possibilities need to be carefully examined and tested if we want to create effective interventions to reduce the toxic type of trolling in online communities and protect netizens.

Finally, the information that is already out there on trolling needs to be translated into bite-sized, practical advice and knowledge for the average netizen, so that we can empower people who find themselves either victims or bystanders of trolling, or those who are tempted to become trolls themselves. In the end, researchers and companies can only do so much to combat toxic trolling; if the users are unable or unwilling to work toward a safer internet themselves, no amount of education or interventions will stop the spread of trolling across cyberspace. Changes in social norms take time (e.g., Heckathorn, 1988), so to combat toxic trolling in the meantime, we need to give tools to the netizens of today to protect themselves from becoming a victim of the trolling cycle. By researching trolling and interventions, and by effectively sharing this knowledge in a clear, understandable way, scholars have the power to make a real impact in the everyday lives of people around the world. For me, there is no higher calling than this.

References

Bandura, A. (1978). Social learning theory of aggression. *Journal of Communication, 28*, 12-29. DOI: 10.1111/j.1460-2466.1978.tb01621.x

Bastiaensens, S., Vandebosch, H., Poels, K., van Cleemput, K., DeSmet, A., & de Bourdeaudhuij, I. (2014). Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. *Computers in Human Behavior, 31*, 259-271. DOI: 10.1016/j.chb.2013.10.036

Bishop, J. (2014). Representations of 'trolls' in mass media communication: A review of media-texts and moral panics relating to 'internet trolling'. *International Journal of Web Based Communities, 10*, 7-24. DOI: 10.1504/IJWBC.2014.058384

Bowman, N. (2018). The demanding nature of video game play. In N.D. Bowman (Ed.), *Video games: A medium that demands our attention* (pp. 1-24). Routledge.

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

Cheng, J., Bernstein, M., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2017, February). *Anyone can become a troll: Causes of trolling behavior in online discussions*. Paper presented at the ACM Conference on Computer Supported Cooperative Work and Social Computing, Portland, Oregon, United States.

Cook, C.L. (2019). Between a troll and a hard place: The demand framework's answer to one of gaming's biggest problems. *Media and Communication, 7*, 176-185. DOI: 10.17645/mac.v7i4.2347

Csikszentmihalyi, M., Abuhamdeh, A., & Nakamura, J. (2005). Flow. In A.J. Elliot & C.S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 598-608). Guildford Publications.

Chapin, J., & Coleman, G. (2017). The cycle of cyberbullying: Some experience required. *The Social Science Journal*, *54*, 314-318. DOI: 10.1016/j.soscij.2017.03.004

Fichman, P, & Sanfilippo, M.R. (2016). *Online trolling and its perpetrators: Under the cyberbridge*. Lanham, MD: Rowman & Littlefield.

Hashimoto, H., & Yamagishi, T. (2013). Two faces of interdependence: Harmony seeking and rejection avoidance. *Asian Journal of Social Psychology*, *16*, 142-151. DOI: 10.1111/ajsp.12022

Heckathorn, D.D. (1988). Collective sanctions and the creation of prisoner's dilemma norms. *American Journal of Sociology*, *94*, 535-562. Retrieved from http://www.jstor.org/stable/2780253

Herring, S., Job-Sluder, K., Scheckler, R., & Barab, S. (2002). Searching for safety online: Managing "trolling" in a feminist forum. *The Information Society*, *18*, 371-384. DOI: 10.1080/01972240290108186

Kwak, H., & Blackburn, J. (2014, October). Linguistic analysis of toxic behavior in an online video game. Paper presented at the first Exploration on Games and Gamers Workshop, Barcelona, Spain.

Paul, H.L., Bowman, N.D., & Banks, J. (2015). The enjoyment of griefing in online games. *Journal of Gaming & Virtual Worlds*, *7*, 243-258. DOI: 10.1386/jgvw.7.3.243_1

Leung, A.K.-Y., & Cohen, D. (2011). Within- and between-culture variation: Individual differences and the cultural logics of honor, face, and dignity cultures. *Journal of Personality and Social Psychology, 100,* 507-526. DOI: 10.1037/a0022151

Markus, H.R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98,* 224-253. DOI: 10.1037/0033-295X.98.2.224

Mora-Cantallops, M., & Sicilia, M.-A. (2018). MOBA games: A literature review. *Entertainment Computing, 26,* 128-138. DOI: 10.1016/j.entcom.2018.02.005

O'Sullivan, P.B., & Flanagin, A.J. (2003). Reconceptualizing 'flaming' and other problematic messages. *New Media & Society, 5,* 69-94. DOI: 10.1177/1461444803005001908

Phillips, W. (2013, March). Ethnography of trolling: Workarounds, discipline jumping & ethical pitfalls (3 of 3). *Ethnography Matters*. Retrieved from https://ethnographymatters.net/blog/2013/03/05/ethnography-of-trolling- workarounds-discipline-jumping-ethical-pitfalls-3-of-3/

Phillips, W. (2011). Meet the trolls. *Index on Censorship, 40,* 68-76. DOI: 10.1177/0306422011409641

Postmes, T., Spears, R., & Lea, M. (1998). Breaching or building social boundaries? SIDE-effects of computer-mediated communication. *Communication Research, 25,* 689-715. DOI: 10.1177/009365098025006006

Sanfilippo, M., Yang, S., & Fichman, P. (2017). Trolling here, there, and everywhere: Perceptions of trolling behaviors in context. *Journal of the Association for Information Science and Technology, 68,* 2313-2327. DOI: 10.1002/asi

Stein, J. (2016, August). Why we're losing the internet to the culture of hate. *Time*. Retrieved from https://time.com/4457110/internet-trolls/

Tedeschi, J.T., Smith III, R.B., & Brown Jr., R.C. (1974). A reinterpretation of research on aggression. *Psychological Bulletin, 81,* 540-562. DOI: 10.1037/h0037028

Thacker, S., & Griffiths, M.D. (2012). An exploratory study of trolling in online video gaming. *International Journal of Cyber Behavior, Psychology and Learning, 2,* 17-33. DOI: 10.4018/ijcbpl.2012100102

Vandebosch, H., & van Cleemput, K. (2009). Cyberbullying among youngsters: Profiles of bullies and victims. *New Media & Society, 11,* 1349-1371. DOI: 10.1177/1461444809341263

Appendices

Appendix 1A

```
#start preprocessing

#Transform to lower case
docs <- tm_map(docs,content_transformer(tolower))

#fix champion names
docs <- tm_map(docs, content_transformer(gsub),
               pattern = paste(char_names, collapse="|"),
               replacement = "charname")

#fix LEETSPEAK (in beginning, end, or middle of a word, as far as possible)
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "(\\<0\\w|\\w0)\\>|\\w0\\w)", replacement = "o")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "(\\<3\\w|\\w3)\\>|\\w3\\w)", replacement = "e")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "(\\<4\\w|\\w4)\\>|\\w4\\w)", replacement = "a")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "(\\<5\\w|\\w5)\\>|\\w5\\w)", replacement = "s")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "(\\<7\\w|\\w7)\\>|\\w7\\w)", replacement = "t")

# fix multiple occuraces of lettres (more than 2 to 2)
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "([[:alnum:]])\\1{2,}", replacement = "\\1\\1")

# fix emoticons (based on wikipedia western emoticons)
# Only ^ - \ ] are special inside character classes.
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "[|>]?[:*;x8=][`'.o^-]?[]dpb3)}>]",
               replacement = " happyface ")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "[|>]?[:*;x8=][`'.o^-]?[[@<c({|]",
               replacement = " sadface ")

#remove (other) punctuation
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "[[:punct:]]", replacement = " ")

#fix GAME ABBREVIATIONS / GAME TERMS
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(ffs|for fucks sake)\\>", replacement =
"for_fuck's_sake")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(ez|ezz|ezzz)\\>", replacement = "easy")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(oom|out of mana)\\>", replacement =
"out_of_mana")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(oop|out of position)\\>", replacement =
"out_of_position")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(ss|switching sides)\\>", replacement =
"switching_sides")
```

```
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(mia|missing in action)\\>", replacement =
"missing_in_action")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(ks|kill steal)\\>", replacement = "kill_steal")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(cc|crowd control)\\>", replacement =
"crowd_control")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(ptfo|play the fucking objective)\\>",
replacement = "play_the_fucking_objective")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(apm|actions per minute)\\>", replacement =
"actions_per_minute")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(brb|be right back)\\>", replacement =
"be_right_back")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(afk|away from keyboard)\\>", replacement =
"away_from_keyboard")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(ggwp|good game well played)\\>", replacement =
"good_game_well_played")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(wp|well played)\\>", replacement =
"well_played")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(gg no re|ggnore|good game no rematch)\\>",
replacement = "good_game_no_rematch")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(gg|good game)\\>", replacement = "good_game")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(glhf|good luck have fun)\\>", replacement =
"good_luck_have_fun")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(bio|bathroom break)\\>", replacement =
"bathroom_break")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(crit|critical strike)\\>", replacement =
"critical_strike")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(qq|quit game|quit|quit the game)\\>",
replacement = "quit_the_game")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(dd|damage dealer)\\>", replacement =
"damage_dealer")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(as|attack speed)\\>", replacement =
"attack_speed")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(dps|damage per second)\\>", replacement =
"damage_per_second")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<ff\\>", replacement = "forfeit")
docs <- tm_map(docs, content_transformer(gsub),
               pattern = "\\<(dmg|damage)\\>", replacement = "damage")
docs <- tm_map(docs, content_transformer(gsub),
```

```
                pattern = "\\<(adc|attack damage carry)\\>", replacement =
"attack_damage_carry")
docs <- tm_map(docs, content_transformer(gsub),
                pattern = "\\<(apc|ability power carry)\\>", replacement =
"ability_power_carry")
docs <- tm_map(docs, content_transformer(gsub),
                pattern = "\\<(ad|attack damage)\\>", replacement =
"attack_damage")
docs <- tm_map(docs, content_transformer(gsub),
                pattern = "\\<(ap|ability power)\\>", replacement =
"ability_power")
docs <- tm_map(docs, content_transformer(gsub),
                pattern = "\\<dmg\\>", replacement = "damage")
docs <- tm_map(docs, content_transformer(gsub),
                pattern = "\\<(inting|intentionally feeding)\\>", replacement
= "itentionally_feeding")
docs <- tm_map(docs, content_transformer(gsub),
                pattern = "\\<(kda|kill death assist ratio)\\>", replacement =
"kill_death_assist_ratio")
docs <- tm_map(docs, content_transformer(gsub),
                pattern = "\\<(bm|bming|bad manners)\\>", replacement =
"bad_manners")

#fix OTHER ABBREVIATIONS
docs <- tm_map(docs, content_transformer(gsub),
                pattern = "\\<srsly\\>", replacement = "seriously")
docs <- tm_map(docs, content_transformer(gsub),
                pattern = "\\<(pls|plz|plse|pleas)\\>", replacement =
"please")
docs <- tm_map(docs, content_transformer(gsub),
                pattern = "\\<(sry|srry)\\>", replacement = "sorry")
docs <- tm_map(docs, content_transformer(gsub),
                pattern = "\\<(np|nop|no prob)\\>", replacement =
"no_problem")

#Strip digits
docs <- tm_map(docs, removeNumbers)

#remove whitespace
docs <- tm_map(docs, stripWhitespace)
```

Appendix 1B

Processing chat data chronemics

From the chat data, both features related to the content and related to the log itself were extracted. The features extracted from the peripheral chat data (non-content related data about messages) were mostly related to the 'high agency' feature from extant literature (see Buckels et al., 2014). Agency here refers to the level of activity demonstrated by the various actors in the trolling interaction, high agency being a high level of activity. High agency in the present study is operationalized as the number of messages sent by each actor.

Processing chat data content

Game chat messages are considered 'messy' in terms of data; they have many spelling errors, game-specific abbreviations, and heavy emoticon usage. This could significantly influence the analysis, especially topic modelling, as this method connects known words semantically (see Blei, Ng, & Jordan, 2003). If the data is full of jargon and misspellings, the algorithms in play will not be able to perform their function properly. Therefore, several data cleaning steps were taken using the topic modelling package 'tm' in R 3.3.3 (Feinerer, Hornik, & Meyer, 2008; Feinerer & Hornik, 2017). All data was cleaned using these automatic techniques.

First, all characters were transformed to lower case. Second, since the chat messages included many different champion names from the game (League of Legends), all champion names were replaced by the placeholder 'Charname'. In the same fashion, all regularly-used emoticons were grouped into 'happyface' and 'sadface,' depending on the expressed sentiment. Because trolling research focuses on positive and negative affect (see Cheng et al., 2017), no further distinction was deemed necessary at this stage. Emoticon processing was done using a publicly-available emoticon database and code (Dasnixon, 2012). Obvious spelling mistakes and

typos were then corrected where possible. The chat messages also included several instances of 'leet-speak,' the use of numbers to refer to letters (e.g., n00b instead of noob). Therefore, leet-speak within words was changed into normal text. Additionally, several (game-related) abbreviations were used. In order to turn this jargon into common English, an open call was placed on the Facebook page of a mid-sized Dutch university's e-sports and gaming student association for lists of gamer abbreviations and terminology with accompanying translations (e.g.,, gg = good game). Members of the association were able to list their responses in the form of comments to the original post. Responses were filtered for relevance to League of Legends and compiled. The abbreviations in the chat data were thus corrected to the full words using the list provided as input for the cleaning code. A complete list of these abbreviations can be found in Appendix C.

Once this initial cleaning was completed, all words were lemmatized using the spaCy parser in Python (Honnibal, 2017). With lemmatization, words are grouped together so they can be analysed as a single term. Based on the meaning of a word in a sentence, the 'lemma' or dictionary form, is determined. For example, 'walking', 'walked', and 'walk' are all changed into 'walk'. The lemmatized texts were created for the opponent messages, teammate messages, troll messages, and all messages combined. At the end of these procedures, the cleaned texts produced were used for all future analyses. With lemmatization complete, we used the textcat package (Hornik et al., 2013) to automatically detect the language of the chat logs. However, due to the high number of slang words and game-related jargon, these were frequently misclassified as languages other than English (e.g.,, Welsh, Scots, etc.) even though manual inspection confirmed that they were indeed English texts. Therefore, we only removed texts that were identified as German, French, and Spanish, as manual inspection confirmed that these were the

most accurately-categorized. This resulted in the removal of 726 messages (out of 8297 messages) from the ‘all’ channel, 667 messages (out of 9091 messages) from the ‘teammates’ channel, and 245 messages (out of 9202 messages) from the ‘opponents’ channel, or approximately 4% of messages total. As a final step, stop words – words with little meaning for topic modelling (e.g., and, the, is) – were removed with the stopword list from the SMART system (Salton, 1971), which is a list of 571 small English words. We performed the same procedure again using R’s built in stopword lists for German, French, and Spanish, as the smaller words from these languages were still showing up in topic models and were not completely removed by the language categorizer.

Additional References

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022. Retrieved from http://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf

Dasnixon (2012). Emoticon module. Retrieved from https://gist.github.com/dasnixon/2969657 (accessed 10 July, 2017).

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software, 25*(5), 1-54. Retrieved from https://www.jstatsoft.org/article/view/v025i05

Feinerer, I., & Hornik, K. (2017). tm. Text Mining Package (R package version 0.7-1) [Computer software]. Retrieved from: https://CRAN.R-project.org/package=tm

Honnibal, M. (2017). SpaCy (version 1.8.2). [Computer software]. Retrieved from: https://spacy.io

Hornik, K., Mair, P., Rauch, J., Geiger, W., Buchta, C., & Feinerer, I. (2013). The textcat package for n-Gram based text categorization in R. *Journal of Statistical Software*, *52*, 1-17. doi: 10.18637/jss.v052.i06

Salton, G. (1971). *The SMART retrieval system: Experiments in automatic document processing*. Englewood Cliffs, NJ: Prentice-Hall.

Appendix 2A

Coding behavioural responses. In addition to their emotions and behavioural intentions, we were also interested in participants' actual behaviour during the game. For this, we saved each individual chat message sent by the participant and coded them according to the schema presented in Table A.

Table A.

Codes given to messages sent by participants in-game.

Code	Code Name	Explanation
1	Rage	A statement or question expressing anger or frustration directed at one of the other players. This usually includes insults to the troll and often punctuation like exclamation marks. If it qualifies as flaming, then it qualifies as rage.
2	Support-seeking	A request for help in dealing with the troll's actions, or a question regarding the game itself. In the case of game mechanics questions, these are general. Otherwise, support-seeking is always a question and is always directed to the other player who is NOT the troll.
3	Negotiation	Attempts to restore harmony to the group. This is done in response to the troll's statements, and is thus always directed at the troll. It usually has a curious or apologetic tone, and is designed to avoid conflict and find a peaceful resolution to the interaction.
4	Argumentation	Refuting the troll's statements. This is similar to negotiation, but its goal is winning a conflict, while negotiation is meant to avoid a conflict, having a discussion instead. Argumentation often has an accusatory tone, while negotiation is more apologetic or curious.
5	Rapport-building	Attempts to build a relationship with the other participants. This is always directed specifically at a player, usually including 'you' or 'your' or some other way to address a specific individual. It is also usually a question.
6	Neutrality	A general statement or question that is not expressing any particular emotion or goal. These are also not directed at a specific player, but rather the whole group generally.
7	Sadness	A statement or question expressing sadness directed at one of the other players. Unlike rage, these should not include any insults to the other player, and should clearly express being hurt or distraught.
8	Playfulness	A statement or question expressing joy or amusement, either general or directed at one of the other players. This should be genuine, indicated by things like exclamation points and possible caps lock. It is also usually not mixed in with obvious

		sarcasm, so if you find sarcasm in the same game, it is rarely if ever genuine playfulness.
9	Sarcasm and Cynicism	A statement or question that uses sarcasm, or expresses boredom or disinterest in the game or the other players' chat. Again, this does not fit in with genuine playfulness – sarcasm trumps playfulness. Use context to determine whether a statement falls in this category or the playfulness category, and also check punctuation – a lack thereof, "…" or a simple "." will usually indicate sarcasm.
10	Other	Anything that does not fall into the previous nine categories.

Two research assistants from each country who were native or near-native speakers of the language in question – either Mandarin, English, or Dutch – gave each individual message sent by our participants one of the codes listed in Table A, knowing the experimental condition. Each two-person team would meet via Skype with the primary investigator at four points during the coding: after 10 participants' messages were coded, then 30 participants', then 75 participants', and a final time when all participants' messages were coded. During these meetings, the coders and primary investigator would discuss messages that were coded differently by the coders and decide upon a final code together. Interrater reliability for this final coding of the last 70+ items was as follows: 81.9% in Taiwan, 87.8% in Pakistan, and 77.3% in the Netherlands.

From there, each code was sorted into one of three macro-categories: retaliation, reparation, and miscellaneous. Codes 1, 4, and 8 – rage, argumentation, and sarcasm and cynicism, respectively – fell under the retaliation category, as all of these possess either an element of disengaging emotions (rage and sarcasm) or conflict initiation. Codes 2, 3, 5, 7, and 8 – support-seeking, negotiation, rapport-building, sadness, and playfulness, respectively – fell under the reparation category, as all of these possess either an engaging emotion (sadness and playfulness) or engage positively with the other players. Codes 6 and 10 – neutrality and other – fell under the miscellaneous category, as they contained no particular emotional content, nor did they engage the other players in any way. It is the two of these three macro-codes – retaliation and reparation – that were used in our analyses.

Appendix 2B

Welcome Message upon Entering Cyberball (English version)

Welcome to Cyberball! During this game, you will be passing a virtual ball to two other players by clicking on their avatars. To chat with them, click the chat box under your avatar, type, and hit enter or click 'send' to send. Your goal is to visualize the game while you play - imagine what the other players look like, where you are playing, the weather, etc.

Don't forget to introduce yourself to the other players too!

Scripts per Condition (English versions)

- **General notes**
  - For all conditions in the Netherlands, Player 3 gives the following introduction after 2 throws: "Hello, I'm Jeroen. I'm from Amsterdam, but study in Tilburg. I enjoy listening to music."
  - For all conditions in Pakistan, Player 3 gives the following introduction after 2 throws: "Hello, I'm Ibrahim. I'm from Istanbul, but study in Lahore. I enjoy listening to music."
  - For all conditions in Taiwan, Player 3 gives the following introduction after 2 throws: "Hello, I'm Jiawei. I'm from Tainan, but study in Taipei. I enjoy listening to music."
- **C1 = In-group control**
  - **Dutch:** Player 1 (computer) says the following script after the first ball toss – "Hi! My name is Thijs. I grew up in here in Tilburg. I'm a big fan of football!" The game then proceeds normally with random ball tosses and no further speech.
  - **Taiwanese:** Player 1 (computer) says the following script after the first ball toss – "Hi! My name is Guanlin. I grew up here in Taipei. I'm a big fan of soccer!" The game then proceeds normally with random ball tosses and no further speech.
  - **Pakistani:** Player 1 (computer) says the following script after the first ball toss – "Hi! My name is Ahmed. I grew up here in Lahore. I'm a big fan of football!" The game then proceeds normally with random ball tosses and no further speech.
- **C2 = In-group flaming**
  - **Dutch:** Player 1 (computer) says the following script after the first ball toss – "Hi! My name is Thijs. I grew up in here in Tilburg. I'm a big fan of football!" Then, every 5 throws, Player 1 says one of the following items:

        - **You're so trashy I can smell you from here.**
        - **Do you not know how to use a computer? Loser.**
        - **I thought you had to be at least 18 to play this game? You play like a child.**
        - **You play like a girl.**
        - **I didn't know you could be so bad at such an easy game.**
    - **Taiwanese:** Player 1 (computer) says the following script after the first ball toss – "Hi! My name is Guanlin. I grew up here in Taipei. I'm a big fan of soccer!" Then, every 5 throws, Player 1 says one of the above items.
    - **Pakistani:** Player 1 (computer) says the following script after the first ball toss – "Hi! My name is Ahmed. I grew up here in Lahore. I'm a big fan of football!" Then, every 5 throws, Player 1 says one of the above items:
- **C3 = In-group ostracism**
    - **Dutch:** Player 1 (computer) says the following script after the first ball toss – "Hi! My name is Thijs. I grew up in here in Tilburg. I'm a big fan of football!" The game then proceeds with neither Players 1 nor 3 (computers) passing to Player 2 (participant) for the duration of the game.
    - **Taiwanese:** Player 1 (computer) says the following script after the first ball toss – "Hi! My name is Guanlin. I grew up here in Taipei. I'm a big fan of soccer!" The game then proceeds with neither Players 1 nor 3 (computers) passing to Player 2 (participant) for the duration of the game.
    - **Pakistani:** Player 1 (computer) says the following script after the first ball toss – "Hi! My name is Ahmed. I grew up here in Lahore. I'm a big fan of football!" The game then proceeds with neither Players 1 nor 3 (computers) passing to Player 2 (participant) for the duration of the game.
- **C4 = Out-group control**
    - **Dutch:** Player 1 (computer) says the following script after the first ball toss – "Hi! My name is Mohammed. I just moved here from Morocco. I'm a big fan of football!" The game then proceeds normally with random ball tosses and no further speech.
    - **Taiwanese:** Player 1 (computer) says the following script after the first ball toss – "Hi! My name is Danilo. I just moved here from the Philippines for work. I'm a big fan of soccer!" The game then proceeds normally with random ball tosses and no further speech.
    - **Pakistani:** Player 1 (computer) says the following script after the first ball toss – "Hi! My name is GulShar. I just moved here from Afghanistan. I'm a big fan of football!" The game then proceeds normally with random ball tosses and no further speech.
- **C5 = Out-group flaming**

    - **Dutch:** Player 1 (computer) says the following script after the first ball toss – "Hi! My name is Mohammed. I just moved here from Morocco. I'm a big fan of football!" Then, every 5 throws, Player 1 says one of the following items:
        - **You're so trashy I can smell you from here.**
        - **Do you not know how to use a computer? Loser.**
        - **I thought you had to be at least 18 to play this game? You play like a child.**
        - **You play like a girl.**
        - **I didn't know you could be so bad at such an easy game.**
    - **Taiwanese:** Player 1 (computer) says the following script after the first ball toss – "Hi! My name is Danilo. I just moved here from the Philippines for work. I'm a big fan of soccer!" Then, every 5 throws, Player 1 says one of the above items.
    - **Pakistani:** Player 1 (computer) says the following script after the first ball toss – "Hi! My name is GulShar. I just moved here from Afghanistan. I'm a big fan of football!" Then, every 5 throws, Player 1 says one of the above items.
- **C6 = Out-group ostracism**
    - **Dutch:** Player 1 (computer) says the following script after the first ball toss – "Hi! My name is Mohammed. I just moved here from Morocco. I'm a big fan of football!" The game then proceeds with neither Players 1 nor 3 (computers) passing to Player 2 (participant) for the duration of the game.
    - **Taiwanese:** Player 1 (computer) says the following script after the first ball toss – "Hi! My name is Danilo. I just moved here from the Philippines for work. I'm a big fan of soccer!" The game then proceeds with neither Players 1 nor 3 (computers) passing to Player 2 (participant) for the duration of the game.

Pakistani: Player 1 (computer) says the following script after the first ball toss – "Hi! My name is GulShar. I just moved here from Afghanistan. I'm a big fan of football!" The game then proceeds with neither Players 1 nor 3 (computers) passing to Player 2 (participant) for the duration of the game.

Appendix 2C

Pre-Test Questionnaire

Please answer the following questions either by filling in the blanks or circling the appropriate response provided.

1. What is your age? _____
2. What is your gender? Male Female Other
3. What is your nationality?
    a. Taiwanese
    b. Pakistani
    c. Dutch
    d. Other _____
4. What is your native language?
    a. Mandarin
    b. Taiwanese
    c. Urdu
    d. Hindi
    e. English
    f. Dutch
    g. Other _____
5. What is your highest completed level of education?
    a. Primary school
    b. Middle school
    c. High school
    d. Vocational training
    e. Bachelor's
    f. Master's
    g. PhD

Below, you will find a number of sentences that describe how you relate to and feel about yourself and others. Read each one and indicate to what extent you feel each sentence is true of you using the scale provided.

1	2	3	4	5
Not at all				Extremely

1. I try hard to work on my reputation (in my relationships with others).
2. I do not consider what others say about me.
3. I wish to have a good reputation.
4. I am rarely concerned about my family's reputation.
5. If my family's reputation is not good, I feel very bad.
6. I find it difficult if others paint an incorrect image of my family.
7. I care about everyone's reputation.
8. I am rarely concerned about others' reputation.
9. I try hard to preserve everyone's reputation (in my interactions with others).

Below, you will find a number of sentences that describe how you relate to others. Read each one and indicate to what extent you feel each sentence is true of you using the scale provided.

1	2	3	4	5
Not at all				Extremely

1. I will sacrifice my self-interest for the benefit of my group.
2. My relationships with others are more important than my personal accomplishments.
3. I will stay in my group if they need me, even when I am not happy with the group.
4. I stick with my group even through difficulties.
5. I try to abide by customs and conventions at university.
6. I help people I know, even if it is inconvenient.
7. I should be judged on my own merit.
8. I am comfortable being singled out for praise and rewards.

Post-Test Questionnaire

This scale consists of a number of words that describe different feelings and emotions. Read each item and then mark the appropriate answer in the space next to that word. Indicate to what extent you currently feel _____; use the following scale to record your answers.

1	2	3	4	5
Not at all				Extremely

_____	Angry	_____	Embarrassed
_____	Happy	_____	Confident
_____	Ashamed	_____	Mad
_____	Proud	_____	Cheerful
_____	Humiliated	_____	Respected

Below, you find a number of statements about various actions you may or may not want to take right now based on your experience playing Cyberball. Please indicate using the scale provided how much you want to …

1	2	3	4	5
Not at all				Extremely

1. … hurt the other players.
2. … have a chat with the other players.
3. … swear at the other players.
4. … meet the other players.
5. … stay away from the other players.
6. … avoid the other players in real life.

	Not at all				*Extremely*
*For each question, please circle the number to the right that best represents the **feelings** you experienced **DURING THE GAME.***					
I felt liked	1	2	3	4	5
I felt rejected	1	2	3	4	5
I felt humiliated	1	2	3	4	5
I felt ridiculed	1	2	3	4	5

Appendix 3A

1. Silence Condition – Neither confederate will speak for the duration of the match.
2. Chat Condition – There will be no troll present – one confederate will say ”gl hf” (good luck, have fun) at the beginning of the match, and will proceed to chat as follows, beginning at the 4 minute mark; all messages are time-stamped):

- I really like ___ as a champ. *(The blank will be filled with the character name of the character they are currently playing.)* [4:00]

- Let’s push top. [5:30]

- What’s going on bot? [7:45]

- Can I get help top? [9:15]

3. Flaming Condition 1– In the pre-game lobby, the troll will immediately say “I call bot.” and choose a champion (character) immediately. At the 4 minute and 45 second mark, the following script begins (if the troll is engaged by the naïve participant, he says ‘whatever, noob’ and continues his tirade against the other confederate; messages are timestamped below):

- Srsly? ___ is such a noob champ. *(The blank will be filled with the character name of the character the other confederate is currently playing.)* [4:45]

- Have you ever even played this game before ___? Fucking retard. *(The blank will be filled with the character name of the character the other confederate is currently playing.)* [6:15]

- ____, uninstall. Fucking easy bots are better than you. *(The blank will be filled with the character name of the character the other confederate is currently playing.)* [7:00]

- Even in fucking experiments I have to carry these noob-ass teams. [8:30]

4. Flaming Condition 2 – In the pre-game lobby, the troll will immediately say “I call bot.” and choose a champion (character) immediately. At the 4 minute mark, the following script begins (script messages are time-stamped below; if the troll is engaged by the naïve participant, he says ‘whatever, noob’ and continues his tirade against the other confederate):

V: I really like ___ as a champ. (*The blank will be filled with the character name of the character they are currently playing.*) [4:00]

T: Srsly? ___ is such a noob champ. *(The blank will be filled with the character name of the character the other confederate is currently playing.)* [4:45]

V: Let’s push top. [5:30]

T: Have you ever even played this game before ___? Fucking retard. *(The blank will be filled with the character name of the character the other confederate is currently playing.)* [6:15]

T: ____, uninstall. Fucking easy bots are better than you. *(The blank will be filled with the character name of the character the other confederate is currently playing.)* [7:00]

V: What's going on bot? [7:45]

T: Even in fucking experiments I have to carry these noob-ass teams. [8:30]

V: Can I get help top? [9:15]

Appendix 3B

Questionnaire 1

Please answer the following questions either by filling in the blanks or circling the appropriate response provided.

6. What is your age? _____
7. With which gender do you identify? Male Female Other
8. What is your nationality? _____
9. What is your country of residence? _____
10. What is your native language? _____
11. What is your estimated level of English proficiency?
    a. Beginner
    b. Intermediate
    c. Advanced
    d. Fluent
    e. Native
12. What is your highest completed level of education?
    a. Primary education
    b. Lower secondary education
    c. Upper secondary education
    d. Bachelor's or equivalent level
    e. Master's or equivalent level
    f. PhD or equivalent level
13. How long have you been playing League of Legends?
    a. > 1 month
    b. 1-6 months
    c. 6 months – 1 year
    d. 1-2 years
    e. 2-4 years
    f. 4-6 years
    g. 6-8 years
    h. 8+ years
14. Which best describes your League of Legends play frequency?
    a. > Once a month
    b. Once a month
    c. Several times a month
    d. Once a week
    e. Several times a week
    f. Daily
15. Do you have a League of Legends account with a competitive rank? Yes No
16. If yes, which rank are you currently? Please select both a rank and division (if applicable).
    a. Bronze a. I
    b. Silver b. II
    c. Gold c. III

d. Platinum d. IV
e. Diamond e. V
f. Master
g. Challenger

17. Do you have experience playing other online multiplayer games? If so, please list a few:

_____
_____
____

Below, you will find a number of words that describe different feelings and emotions. Read each word and indicate to what extent you feel this way right now, that is, at the present moment. You can mark your appropriate answer in the space next to that word, using the following scale:

1 2 3 4 5
Not at all Extremely

_____	Interested	_____	Irritable
_____	Distressed	_____	Alert
_____	Excited	_____	Ashamed
_____	Upset	_____	Inspired
_____	Strong	_____	Nervous
_____	Guilty	_____	Determined
_____	Scared	_____	Attentive
_____	Hostile	_____	Jittery
_____	Enthusiastic	_____	Active
_____	Proud	_____	Afraid

<u>Questionnaire 2</u>

This scale consists of a number of words that describe different feelings and emotions. Read each item and then mark the appropriate answer in the space next to that word. Indicate to what extent you felt _____ during the game; use the following scale to record your answers.

1 2 3 4 5
Not at all Extremely

_____	Interested	_____	Irritable
_____	Distressed	_____	Alert
_____	Excited	_____	Ashamed
_____	Upset	_____	Inspired
_____	Strong	_____	Nervous
_____	Guilty	_____	Determined
_____	Scared	_____	Attentive
_____	Hostile	_____	Jittery
_____	Enthusiastic	_____	Active
_____	Proud	_____	Afraid

Below, you find a number of statements about the other players in the game you just played. Please rate the statements below using the following scale:

1	2	3	4	5
Strongly disagree	Disagree	Neutral	Agree	Strongly agree

Based on the game we just played, I think that Player 1 is …

___ Friendly___ Kind

___ Intelligent ___ Knowledgeable

___ Creative ___ Incompetent

___ Mean ___ Trustworthy

Based on the game we just played, I think that Player 2 is …

___ Friendly___ Kind

___ Intelligent ___ Knowledgeable

___ Creative ___ Incompetent

___ Mean ___ Trustworthy

Read each statement carefully and as you answer the questions think of your team during the game as a whole. For each statement fill in the box under the MOST APPROPRIATE heading that best describes the group during the game. Please mark only ONE box for each statement.

	Not at all	A little bit	Moderately	Quite a bit	Extremely
There was a sense of team spirit.					
The team got along well together.					
There was friction and anger between the members.					
The members affronted each other.					
The members rejected each other.					
The members distrusted each other.					
The members felt comfortable with one another.					
The members appeared tense.					

Please answer the following questions as completely and truthfully as possible.

1. Please rate your team's level of teamwork on a scale of 1 to 10, with 1 being non-existent and 10 being a well-oiled machine. Justify your answer below.

Rating: 1 2 3 4 5 6 7 8 9 10

Justification: _____
_____
_____
_____
_____

2. How did you feel during the game? Describe your experience.

_____
_____
_____
_____
_____

3. Would you have reported (i.e., sent a complaint and request for punishment to the game authorities) any other players were this a real game? Why or why not?

_____
_____
_____
_____
_____

Summary

Within the world of online gaming, trolling has become a regular menace. While gamers try to connect and socialize with one another, or even simply play the game, there are other gamers – trolls – on the prowl for an entirely different kind of good time, one in which they are enjoying themselves at the expense of everyone else (Chapters 2 and 3). Although trolling is common, and mass-media has latched onto it as a hot topic, it is only recently that the academic community has begun to take a serious look at how trolling occurs in and affects the gaming community at large. However, a lot of this literature is either descriptive in nature (see Thacker & Griffiths, 2012), or jumps ahead to prevention (see Cheng et al., 2017) without taking a deeper look at more than a single underlying motivation at a time. In short, there is a complex and prolific phenomenon happening online, but the research on it is only emerging.

This dissertation's goal is to take a deeper look at trolling as a phenomenon, beyond what has been done so far. More specifically, I aim to figure out a) what trolling is, b) why people do it, and c) who helps and who hinders trolling in online games. To do this, I took four different perspectives: the troll's (Chapter 2), the researcher's (Chapter 3), the victim's (Chapter 4), and the bystander's (Chapter 5). The purpose of Chapter 2 is to give the troll's perspective on trolling, something that researchers had yet to do at the time. To do this, I interviewed 22 people who said that they had a history of trolling in online games. More specifically, I asked them about times they witnessed, were victims of, or perpetrated trolling, as well as what they thought about how the gaming community dealt with and felt about trolls and trolling. My goal with these interviews was threefold: I wanted to figure out a) what trolls consider trolling, b) what motivates them to do it, and c) the role of everyone else in game when it comes to encouraging or discouraging more trolling. What I found was that although trolling was almost universally

considered a negative part of online gaming culture, and all the trolls in our group of participants started as victims of trolls before becoming trolls themselves, the online community neither encourages nor discourages it, making it an asocial activity.

The next chapter allowed me to look at an archive of trolling incidents to find patterns in the way that different people involved in real-life trolling incidents communicate with one another. This public online archive consisted of 10,000 reported incidents of trolling in the popular online game League of Legends, and it included game data like player statistics, as well as everything all the players involved said during the game. Once the data was properly cleaned and prepared, myself and my co-author, Dr. Rianne Conijn, analysed the chat logs in two different ways: structural topic modelling (STM), and a traditional dictionary-based content analysis. In this way, we were able to see what characterized all the different actors – the troll, their victim(s), and the bystanders – and what was similar when it came to their messages. All this information was then compared to what existed already in literature used to describe trolls and trolling and complement what I had learned about trolls from Chapter 2. The key finding was that trolls and their teammates actually share a lot of the negative speech patterns (e.g., profanity, negative emotional content) normally associated with only trolls. Practically, this means that we have to be extremely careful as researchers when labelling trolls for the purpose of study, as we could very easily be falsely labelling victims.

After speaking to trolls and looking at trolling interactions broadly, Chapter 4 focuses intently on the victim and their personal experience in a trolling simulation, taking into account their cultural background and values. It is also the first study to directly compare and contrast two different types of trolling: verbal (flaming) and behavioural (ostracism). They are both really common online occurrences, so the participants could easily relate, but they are extremely

different in how they are executed, with flaming being vicious insults and ostracism being totally ignoring a person. Our participants were either Dutch, Pakistani, or Taiwanese, so that we could also look at how people from vastly different cultural backgrounds would react to – behaviourally and emotionally – the different kinds of trolling in the study. We simulated a trolling experience by putting our participants in a virtual game of catch with two computerized co-players, who they were led to believe were real people of either the same nationality or a minority member (e.g., a Moroccan immigrant in the Netherlands), who I had programmed to either troll them or silently watch the trolling happen. We found that there are indeed cultural differences when it comes to reactions, as well as differences between reactions to the two trolling types, but the core take-away is that future trolling interventions have to take into account the cultures of the target population as well as the specific type of trolling they are trying to fix or prevent in order to be effective.

In the penultimate chapter, I shift the focus one last time to bystanders by putting participants in a game of League of Legends with two confederates who would troll one another throughout the game. This study's goal was to see what motivated gamers to report trolls to an authority figure (the game developer) using the game's built-in reporting functions, as the results of Chapter 2's study suggested that this was an effective trolling deterrent. It is also, according to the results of the same study, the least-used recourse by bystanders faced with trolls in the proverbial wild. We found that how warm and friendly the troll was perceived to be and how competent the victim was perceived to be were what determined whether the participant reported our fake troll or not. A more competent victim and a less warm troll lead to more reports.

To conclude, there is still a lot more to learn about trolls and trolling, but the field is farther along now than when this project started in 2015. There is a broad definition developed

that encompasses most of the descriptive literature on trolling in games thus far. We also now know that there is the indication of a trolling cycle that requires further exploration. This is particularly important to know when it comes to the world of game development, as knowing the cycle exists allows for multiple points of intervention in order to protect their customers. Finally, this dissertation has shown the complexity of not just trolls – who are often portrayed in the media as one-dimensional antagonists – but also of everyone else involved in trolling interactions. Trolls, victims, and bystanders are all multi-faceted humans, and trolling, like all interactions, is an intricate social dance that deserves to be studied in even further depth in the future than what I have done here.

List of Publications

Cook, C.L. (2019). Between a troll and a hard place: The demand framework's answer to one of gaming's biggest problems. Media and Communication: Video Games as Demanding Technologies, 7(4). DOI: 10.17645/mac.v7i4.2347

Cook, C., Conijn, R., Antheunis, M., & Schaafsma, J. (2019). For whom the gamer trolls: A study of trolling interactions in the online gaming context. Journal of Computer-Mediated Communication, 24(6), 293-318. DOI: 10.1093/jcmc/zmz014

Cook, C., Schaafsma, J., & Antheunis, M. (2018). Under the bridge: An in-depth examination of online trolling in the gaming context. New Media & Society, 20(9), 3323-3340. DOI: 10.1177/1461444817748578

LaMastra, N., Uttarapong, J., Gandhi, R., Cook, C.L., & Wohn, D.Y. (2020, November). How a live streamer's choice in played game affects mental health conversations. In CHI PLAY '20: Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play (pp. 297-300). DOI: 10.1145/3383668.3419894

Verheij, T., Bleize, D., & Cook, C. (2020). Friendly fire off: Does cooperative gaming in a competitive setting lead to prosocial behaviour? Press Start, 6(1). Retrieved from https://press-start.gla.ac.uk/index.php/press-start/article/view/185

Voyer, D., Saint-Aubin, J., & Cook, C. (2014). Strategies and pseudoneglect on luminance judgments: An eye-tracking investigation. Journal of Experimental Psychology: Human Perception and Performance, 40(5), 1789-1798. DOI: 10.1037/a0037790

TiCC PhD Series

1. Pashiera Barkhuysen. *Audiovisual Prosody in Interaction.* Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 3 October 2008.
2. Ben Torben-Nielsen. *Dendritic Morphology: Function Shapes Structure.* Promotores: H.J. van den Herik, E.O. Postma. Co-promotor: K.P. Tuyls. Tilburg, 3 December 2008.
3. Hans Stol. *A Framework for Evidence-based Policy Making Using IT.* Promotor: H.J. van den Herik. Tilburg, 21 January 2009.
4. Jeroen Geertzen. *Dialogue Act Recognition and Prediction.* Promotor: H. Bunt. Co-promotor: J.M.B. Terken. Tilburg, 11 February 2009.
5. Sander Canisius. *Structured Prediction for Natural Language Processing.* Promotores: A.P.J. van den Bosch, W. Daelemans. Tilburg, 13 February 2009.
6. Fritz Reul. *New Architectures in Computer Chess.* Promotor: H.J. van den Herik. Co-promotor: J.W.H.M. Uiterwijk. Tilburg, 17 June 2009.
7. Laurens van der Maaten. *Feature Extraction from Visual Data.* Promotores: E.O. Postma, H.J. van den Herik. Co-promotor: A.G. Lange. Tilburg, 23 June 2009 (cum laude).
8. Stephan Raaijmakers. *Multinomial Language Learning.* Promotores: W. Daelemans, A.P.J. van den Bosch. Tilburg, 1 December 2009.
9. Igor Berezhnoy. *Digital Analysis of Paintings.* Promotores: E.O. Postma, H.J. van den Herik. Tilburg, 7 December 2009.
10. Toine Bogers. *Recommender Systems for Social Bookmarking.* Promotor: A.P.J. van den Bosch. Tilburg, 8 December 2009.
11. Sander Bakkes. *Rapid Adaptation of Video Game AI.* Promotor: H.J. van den Herik. Co-promotor: P. Spronck. Tilburg, 3 March 2010.

12. Maria Mos. *Complex Lexical Items.* Promotor: A.P.J. van den Bosch. Co-promotores: A. Vermeer, A. Backus. Tilburg, 12 May 2010 (in collaboration with the Department of Language and Culture Studies).
13. Marieke van Erp. *Accessing Natural History. Discoveries in data cleaning, structuring, and retrieval*. Promotor: A.P.J. van den Bosch. Co-promotor: P.K. Lendvai. Tilburg, 30 June 2010.
14. Edwin Commandeur. *Implicit Causality and Implicit Consequentiality in Language Comprehension*. Promotores: L.G.M. Noordman, W. Vonk. Co-promotor: R. Cozijn. Tilburg, 30 June 2010.
15. Bart Bogaert. *Cloud Content Contention*. Promotores: H.J. van den Herik, E.O. Postma. Tilburg, 30 March 2011.
16. Xiaoyu Mao. *Airport under Control.* Promotores: H.J. van den Herik, E.O. Postma. Co-promotores: N. Roos, A. Salden. Tilburg, 25 May 2011.
17. Olga Petukhova. *Multidimensional Dialogue Modelling*. Promotor: H. Bunt. Tilburg, 1 September 2011.
18. Lisette Mol. *Language in the Hands.* Promotores: E.J. Krahmer, A.A. Maes, M.G.J. Swerts. Tilburg, 7 November 2011 (cum laude).
19. Herman Stehouwer. *Statistical Language Models for Alternative Sequence Selection*. Promotores: A.P.J. van den Bosch, H.J. van den Herik. Co-promotor: M.M. van Zaanen. Tilburg, 7 December 2011.
20. Terry Kakeeto-Aelen. *Relationship Marketing for SMEs in Uganda*. Promotores: J. Chr. van Dalen, H.J. van den Herik. Co-promotor: B.A. Van de Walle. Tilburg, 1 February 2012.

21. Suleman Shahid. *Fun & Face: Exploring Non-Verbal Expressions of Emotion during Playful Interactions*. Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 25 May 2012.
22. Thijs Vis. *Intelligence, Politie en Veiligheidsdienst: Verenigbare Grootheden?* Promotores: T.A. de Roos, H.J. van den Herik, A.C.M. Spapens. Tilburg, 6 June 2012 (in collaboration with the Tilburg School of Law).
23. Nancy Pascall. *Engendering Technology Empowering Women.* Promotores: H.J. van den Herik, M. Diocaretz. Tilburg, 19 November 2012.
24. Agus Gunawan. *Information Access for SMEs in Indonesia*. Promotor: H.J. van den Herik. Co-promotores: M. Wahdan, B.A. Van de Walle. Tilburg, 19 December 2012.
25. Giel van Lankveld. *Quantifying Individual Player Differences.* Promotores: H.J. van den Herik, A.R. Arntz. Co-promotor: P. Spronck. Tilburg, 27 February 2013.
26. Sander Wubben. *Text-to-text Generation Using Monolingual Machine Translation.* Promotores: E.J. Krahmer, A.P.J. van den Bosch, H. Bunt. Tilburg, 5 June 2013.
27. Jeroen Janssens. *Outlier Selection and One-Class Classification*. Promotores: E.O. Postma, H.J. van den Herik. Tilburg, 11 June 2013.
28. Martijn Balsters. *Expression and Perception of Emotions: The Case of Depression, Sadness and Fear*. Promotores: E.J. Krahmer, M.G.J. Swerts, A.J.J.M. Vingerhoets. Tilburg, 25 June 2013.
29. Lisanne van Weelden. *Metaphor in Good Shape.* Promotor: A.A. Maes. Co-promotor: J. Schilperoord. Tilburg, 28 June 2013.
30. Ruud Koolen. *"Need I say More? On Overspecification in Definite Reference."* Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 20 September 2013.

31. J. Douglas Mastin. *Exploring Infant Engagement. Language Socialization and Vocabulary. Development: A Study of Rural and Urban Communities in Mozambique.* Promotor: A.A. Maes. Co-promotor: P.A. Vogt. Tilburg, 11 October 2013.
32. Philip C. Jackson. Jr. *Toward Human-Level Artificial Intelligence – Representation and Computation of Meaning in Natural Language.* Promotores: H.C. Bunt, W.P.M. Daelemans. Tilburg, 22 April 2014.
33. Jorrig Vogels. *Referential Choices in Language Production: The Role of Accessibility.* Promotores: A.A. Maes, E.J. Krahmer. Tilburg, 23 April 2014.
34. Peter de Kock. *Anticipating Criminal Behaviour.* Promotores: H.J. van den Herik, J.C. Scholtes. Co-promotor: P. Spronck. Tilburg, 10 September 2014.
35. Constantijn Kaland. *Prosodic Marking of Semantic Contrasts: Do Speakers Adapt to Addressees?* Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 1 October 2014.
36. Jasmina Marić. *Web Communities, Immigration and Social Capital.* Promotor: H.J. van den Herik. Co-promotores: R. Cozijn, M. Spotti. Tilburg, 18 November 2014.
37. Pauline Meesters. *Intelligent Blauw*. Promotores: H.J. van den Herik, T.A. de Roos. Tilburg, 1 December 2014.
38. Mandy Visser. *Better Use Your Head. How People Learn to Signal Emotions in Social Contexts.* Promotores: M.G.J. Swerts, E.J. Krahmer. Tilburg, 10 June 2015.
39. Sterling Hutchinson. *How Symbolic and Embodied Representations Work in Concert.* Promotores: M.M. Louwerse, E.O. Postma. Tilburg, 30 June 2015.
40. Marieke Hoetjes. *Talking hands. Reference in Speech, Gesture and Sign.* Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 7 October 2015

41. Elisabeth Lubinga. *Stop HIV. Start Talking? The Effects of Rhetorical Figures in Health Messages on Conversations among South African Adolescents*. Promotores: A.A. Maes, C.J.M. Jansen. Tilburg, 16 October 2015.
42. Janet Bagorogoza. *Knowledge Management and High Performance. The Uganda Financial Institutions Models for HPO.* Promotores: H.J. van den Herik, B. van der Walle, Tilburg, 24 November 2015.
43. Hans Westerbeek. *Visual realism: Exploring Effects on Memory, Language Production, Comprehension, and Preference.* Promotores: A.A. Maes, M.G.J. Swerts. Co-promotor: M.A.A. van Amelsvoort. Tilburg, 10 February 2016.
44. Matje van de Camp. *A link to the Past: Constructing Historical Social Networks from Unstructured Data.* Promotores: A.P.J. van den Bosch, E.O. Postma. Tilburg, 2 March 2016.
45. Annemarie Quispel. *Data for all: Data for all: How Professionals and Non-Professionals in Design Use and Evaluate Information Visualizations.* Promotor: A.A. Maes. Co-promotor: J. Schilperoord. Tilburg, 15 June 2016.
46. Rick Tillman. *Language Matters: The Influence of Language and Language Use on Cognition.* Promotores: M.M. Louwerse, E.O. Postma. Tilburg, 30 June 2016.
47. Ruud Mattheij. *The Eyes Have It*. Promoteres: E.O. Postma, H. J. Van den Herik, and P.H.M. Spronck. Tilburg, 5 October 2016.
48. Marten Pijl, *Tracking of Human Motion over Time.* Promotores: E. H. L. Aarts, M. M. Louwerse. Co-promotor: J. H. M. Korst. Tilburg, 14 December 2016.

49. Yevgen Matusevych, *Learning Constructions from Bilingual Exposure: Computational Studies of Argument Structure Acquisition*. Promotor: A.M. Backus. Co-promotor: A.Alishahi. Tilburg, 19 December 2016.
50. Karin van Nispen. *What Can People with Aphasia Communicate with their Hands? A Study of Representation Techniques in Pantomime and Co-Speech Gesture*. Promotor: E.J. Krahmer. Co-promotor: M. van de Sandt-Koenderman. Tilburg, 19 December 2016.
51. Adriana Baltaretu. *Speaking of Landmarks. How Visual Information Inuences Reference in Spatial Domains.* Promotores: A.A. Maes and E.J. Krahmer. Tilburg, 22 December 2016.
52. Mohamed Abbadi. *Casanova 2, a Domain Specific Language for General Game Development.* Promotores: A.A. Maes, P.H.M. Spronck and A. Cortesi. Co-promotor: G. Maggiore. Tilburg, 10 March 2017.
53. Shoshannah Tekofsky. *You Are Who You Play You Are. Modelling Player Traits from Video Game Behavior.* Promotores: E.O. Postma and P.H.M. Spronck. Tilburg, 19 June 2017.
54. Adel Alhuraibi, *From IT-BusinessStrategic Alignment to Performance: A Moderated Mediation Model of Social Innovation, and Enterprise Governance of IT*. Promotores: H.J. van den Herik and Prof. dr. B.A. Van de Walle. Co-promotor: Dr. S. Ankolekar. Tilburg, 26 September 2017.
55. Wilma Latuny. *The Power of Facial Expressions*. Promotores: E.O. Postma and H.J. van den Herik. Tilburg, 29 September 2017.

56. Sylvia Huwaë, *Different Cultures, Different Selves? Suppression of Emotions and Reactions to Transgressions across Cultures*. Promotores: E.J. Krahmer and J. Schaafsma. Tilburg, 11 October, 2017.
57. Mariana Serras Pereira, *A Multimodal Approach to Children's Deceptive Behavior*. Promotor: M. Swerts. Co-promotor: S. Shahid Tilburg, 10 January, 2018.
58. Emmelyn Croes, *Meeting Face-to-Face Online: The Effects of Video-Mediated Communication on Relationship Formation.* Promotores: E.J. Krahmer and M. Antheunis. Co-promotor: A.P. Schouten. Tilburg, 28 March 2018.
59. Lieke van Maastricht, *Second language prosody: Intonation and rhythm in production and perception*. Promotores: E.J. Krahmer, M.G.J. Swerts. Tilburg, 9 May 2018.
60. Nanne van Noord, *Learning visual representations of style*. Promotores: E.O. Postma, M. Louwerse. Tilburg, 16 May 2018.
61. Ingrid Masson Carro, *Handmade: On the cognitive origins of gestural representations*. Promotor: E.J. Krahmer. Co-promotor: M.B. Goudbeek. Tilburg, 25 June 2018.
62. Bart Joosten, *Detecting social signals with spatiotemporal Gabor filters*. Promotores: E.J. Krahmer, E.O. Postma. Tilburg, 29 June 2018
63. Yan Gu, Chinese hands of time: *The effects of language and culture on temporal gestures and spatio-temporal reasoning*. Promotor: M.G.J. Swerts. Co-promotores: M.W. Hoetjes, R. Cozijn. Tilburg, 5 June 2018.
64. Thiago Castro Ferreira, *Advances in natural language generation: Generating varied outputs from semantic inputs*. Promotor: E.J. Krahmer. Co-promotor: S. Wubben. Tilburg, 19 September 2018.

65. Yu Gu, *Automatic emotion recognition from Mandarin speech*. Promotores: E.O. Postma, H.J. van den Herik, H.X. Lin. Tilburg, 28 November 2018.
66. Francesco Di Giacomo, *Metacasanova: A high-performance meta-compiler for domain-specific languages*. Promotores: P.H.M Spronck, A. Cortesi, E.O. Postma. Tilburg, 19 November 2018.
67. Ákos Kádár, *Learning visually grounded and multilingual representations*. Promotores: E.O. Postma, A. Alishahi. Co-promotor: G.A. Chrupala. Tilburg, 13 November 2019.
68. Phoebe Mui, *The many faces of smiling: Social and cultural factors in the display and perception of smiles*. Promotor: M.G.J. Swerts. Co-promotor: M.B. Goudbeek. Tilburg, 18 December 2019.
69. Véronique Verhagen, *Illuminating variation: Individual differences in entrenchment of multi-word units.* Promotor: A.M. Backus. Co-promotores: M.B.J. Mos, J. Schilperoord. Tilburg, 10 January 2020.
70. Alain Hong, *Women in the Lead: Gender, Leadership Emergence, and Negotiation Behavior from a Social Role Perspective*. Promotor: J. Schaafsma. Co-promotor: P.J. van der Wijst. Tilburg, 3 June 2020.
71. Debby Damen, *Taking perspective in communication: Exploring what it takes to change perspectives*. Promotor: E.J. Krahmer. Co-promotores: M.A.A. Van Amelsvoort, P.J. Van der Wijst. Tilburg, 4 November 2020.

Acknowledgements

I will start this section with a confession, as I know that this is the part of the thesis that everyone will actually read. I am writing this when I should be revising not one, but three articles. I remember within the first few weeks of starting this whole process, I ended up getting carted off to the doctoral defense of a woman I had never met, and the only part of her thesis I read was her acknowledgements. If you are that student in my own defense, you should know that I am currently a ball of terror and stress, but you are reading this, which means I made it through and finally finished in the end. Of course, I did not get through alone, hence this section of the last four-ish years of my life.

You have a long defense to sit through, dear reader, so let's start with the thank-yous. I'll start with the person who started my PhD journey: Ms. Sally Chan, the fantastic postdoc who used to work in my lab when I was still a master's student back in Cambridge, UK. Another fun story for you reader – I was not supposed to do my PhD here at Tilburg University. I had been accepted elsewhere in Psychology, but my funding fell through about 3 months before I was set to start. When this happened, my entire lab, but especially Sally, banded together to find me a new program with funding. It was Sally that found me the program I graduate from today. To this day, I'm sure you could ask Sally for a stick of uranium and a chinchilla and she'd either have it in her backpack or be able to get it for you within a day. Thank you, Sally – may we both have a long life of financial security and scientific discovery.

While we're (metaphorically) still in Cambridge, I'd like to send some extra thank-yous to the rest of my old lab members and reminisce. To Matt, I will never be able to separate you from multi-level modelling. I never got to do it myself, but whenever I see it in a paper (especially if that paper's out of 'Straya), I always think of you. To Sai, you introduced me to

hotpot AND took me to China! Every Christmas now, I always think back to having dinner with your parents around those little heaters while all the heat went out the open windows. It's still one of my favourite trips, and I wish you all the best with your social media career! To Antonia, I fully expect you to be running the world sometime soon, so please remember me when it happens. You've always been absurdly confident when it comes to my abilities, so right back at you, lady. To Rui, I've been lucky enough to see you again after I left Cambridge, but I still hope that I can grow up to be as strong and elegant as you someday. May you always be the 老虎妈妈 we all aspire to be, deep down. To Laurie and Lara, because I always think of you as a pair, I hope all your yoga dreams come true! You always come to mind whenever I have a miso soup cleanse or see just about anything involving clean eating. To Maurice, the man who taught me what a "thirst trap" was, may you be the godliest of all doctors, and may all your paths be blessed. I look forward to seeing you bring all your passions to the world! To the rest of the postdoc army – Bryant, Joe, and Laura – thank you for being honest about your postdoc experience, letting us see all the highs and lows and showing me just how hard you have to work to make it. You all deserve all the success in the world, and I will be thoroughly disappointed in the world if it doesn't happen, but proud of you all regardless.

Of course, I can't forget my Newnham girls either. Catherine, Lord of the Zombie Cells – Yi Hui, goddess of the Singaporean Education system – and Angela, the Canadian connection I so desperately needed abroad – I hope we never have to resort to starting an international underground Mah Zhong gambling ring to support ourselves, but I am forever thankful that we totally could if the need arose. To the rowing team, you gave me my first real taste of successful athleticism in my life. Jenna in particular, you're the single greatest coach I've ever had in … anything, really. Thank you all for showing me that I am not a brain on a stick, but rather a stick

that can grow muscles and navigate a river on a boat. To the MCR ladies, I made it out of Cambridge and still got a PhD! I think that deserves a round of Elderflower Cordial, on me.

On va prendre une petite pause ici pour remercier ceux qui m'ont encouragée même avant que j'ai déménagé à l'autre côté de l'océan. Je veux étendre un grand merci pour mes amies des labos de cognition; vous êtes tous des reines (oui, Sébass inclu!). J'espère qu'un jour on pourra tous prendre des quiz stupides ensembles encore. Je veux aussi dire merci aux professeurs qui m'ont supervisée lorsque j'étais à Moncton: Jean, François, Geneviève, et Kathérine. Vous m'avez écris tant de lettres de référence, et j'ai hâte à faire la même chose pour mes propres étudiants. Un note finale pour deux dames spéciales – Jézabel et Myriam – j'espère qu'on pourra se revoir bientôt! Vous me manquez, toutes les deux.

After that brief séjour in the Acadian section of Eastern Canada, let's head to where we are today: The Netherlands. Of course, I have to send along a general thank-you to both halves of the original Communications department that hired me; most everyone in both has had at the very least a small hand in today, but there are certain people who need some extra special mentions. To my original officemates – Karin and Lieke, neither of whom are in Tilburg any longer – I wish you all the success in the world with your little families and big careers. To my ICA compatriots, professors and students alike, I hope to keep seeing you every year as we all go forward. To Emmelyn, my academic big sister, keep looking for enthusiastic nods at all your presentations. And to my most recent officemates – Juliette's ERC crew: Marlies, Marieke, and Thia – I'll keep liking your tweets as long as the project goes on, and will think of you whenever I hear an apology (which, as a Canadian, is basically on a minutely basis). I also can't forget Emiel, for letting me call from a random man's phone in China to accept this job, and accepting that as proof of life/existence. Of course, the most special of academic thank-yous go to Juliette

and Marjolijn, my supervisors. I hope that your future PhD students are major improvements over this hot mess on (virtual) stage. This book and the individual publications within are shaped by your guidance, and are a testament to your skill and experience. Let's not forget my committee either – an international group of fabulous scholars I am happy to join today!

Finally, we make it to the non-academic thank-yous. First and foremost, my family, with my parents getting a special mention for coming overseas to pick up the pieces of my broken mind. It has probably been an even longer four years for you than me watching from across the ocean. Hopefully if I end up outside of Canada, I move somewhere exciting to visit, and if I end up back home, I can take full charge of holiday turkeys on a permanent basis. To my paranimfen and Dutch unofficial family members – Rianne, Jan-Willem, Emilie, Naomi, and Maurits – I'll have you know that I wish with all my heart that I could be there in person today, for the sole reason of hanging out again and playing board games. They could be giving me 10 PhD's today and it wouldn't live up to the hype of seeing even one of you. Thank you all for being the best part of the last four (ish) years, and at least in my top 5 best parts of living for the rest of my life.

And Chuck. Poor, long-suffering Chuck – the only human being who has, by choice, not only put up with me for over a decade, but actively listened and supported me on a daily basis when I was at my least livable. You deserve better, but you're literally stuck with me. Forever. You have all of my love, and even more of my sympathy. You poor, poor man. God help you.

Which segues nicely into the dedication of this work, summed up in the following three words:

Soli Deo Gloria.