Why do People Engage in Online Shaming?

A Social Identity Approach

By

Kuni Zhao

BPsyc (Hons)

Thesis

Submitted to Flinders University

for the degree of

Doctor of Philosophy (Psychology)

College of Education, Psychology and Social Work

Wednesday 22nd of October 2025

Table of Contents

Summary

In recent years, the practice of online shaming, or calling out individuals on social media, has become increasingly prevalent. Public and media attention surrounding this phenomenon has also increased – perhaps especially on the detrimental consequences that online shaming can cause for its target. Various terms were coined to describe this phenomenon, such as internet shaming, online public outrage, online vigilantism, social media fireworks, "going viral", and in the Chinese context, "human flesh search". However, in the academic literature about online shaming, there is little consensus on the motivations behind this phenomenon, or even on how online shaming should be conceptualised. To address these gaps, my research examines how we can conceptualise this phenomenon and seeks to understand why people engage in online shaming.

Through analysing the multidisciplinary literature on online shaming, I find that this behaviour frequently has a clear collective, or group-based, dimension: online shaming often manifests as a collective behaviour in which likeminded individuals appear to pursue the same goal(s). Based on these propositions, I take a social identity approach and argue that online shaming can be influenced by shared social identities. Specifically, I argue that there is a need to systematically and empirically examine online shaming as a group behaviour that involves both group processes and the pursuit of shared (group-level) goals. I further argue that online shaming involves a variety of motives: it can be shaped by the motivation to inflict harm and punishment on people perceived as having transgressed; however, it can also reflect a justice motive as well as a commitment to one's social group.

Across four studies, I use different methods to examine the goal(s) that drive people's online shaming behaviour, as well as test my propositions about the group processes that influence people's engagement. Chapter 2 reports the findings of a scoping and narrative review that systematically examines the multidisciplinary literature on online shaming. The

review presents a thematic analysis that integrates the online shaming literature into several overarching psychological goals. Specifically, the identified psychological goals are 1) punishing the perceived wrongdoer, 2) deterring the perceived wrongdoing, 3) seeking social acknowledgement, and 4) creating change. These identified goals formed a basis for my subsequent studies.

In Chapter 3, I further investigate the goals through analysing a real-life online shaming event on Twitter (Study 2). With the use of topic modelling that involves natural language processing techniques, Study 2 findings support for the goals identified in Chapter 2. Specifically, in this "real world" instance of online shaming, the analysis reveals evidence for punishment, deterrence, and social change goals. Moreover, the analysis also reveals that the events were explainable by group processes: distinctive groups emerged during online shaming, and that online shaming was shaped by both intragroup and intergroup interactions.

In Chapter 4, I present two experimental studies to examine whether, and how, group processes influence people's online shaming engagement. I approach the questions from a lens of leadership-followership dynamic, using a paradigm inspired by the Milgram's and examining whether people can be mobilised by social identity leaders to engage in online shaming. The results reveal that people generally show resistance to shaming punishments, irrespective of whether they share an identity with the leader. Nonetheless, when considering online shaming in specific situations, the leadership-followership dynamic can still shape people's shaming engagements.

Overall, the findings suggest that online shaming is characterised by group dynamics and motivated by multiple discrete goals. The identified goals suggest that online shaming is not just a negative behaviour that people engage to harm or "bring shame upon" – rather, people engage in online shaming for motives relating to punishment, deterrence, but also social acknowledgement, and to create social change. Underlying these goals is the idea that

online shaming can serve to fulfill one's identity-based needs, which may potentially benefit the ingroup and/or society at large. One implication is that, in some cases, online shaming can be analogous to online activism. In other cases, however, the online shaming punishments can lead to reluctance and reactance, suggesting that it is a practice that needs to be understood in context. Insights from this thesis may contribute to both theoretical understanding and practical implications of online shaming, informing social actors such as the press, social media platforms, and policymakers to better address online shaming.

Declaration

I certify that this thesis:

1. does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university

2. and the research within will not be submitted for any other future degree or diploma without the permission of Flinders University; and

3. to the best of my knowledge and belief, does not contain any material previously published or written by another person except where due reference is made in the text; and

4. a professional editor was not used in the preparation of this thesis. However, an Artificial Intelligence platform (ChatGPT by OpenAI) was used for editing process and to identify improvements in writing style, in accordance with the current guidance at Flinders University.

Signed.....**Zijun Zhao (Kuni)**.....

Date..... **Friday 16th of May 2025**.....

Acknowledgements

I gratefully acknowledge the support of an Australian Government Research Training Program (RTP) Scholarship for this research (doi.org/10.82133/C42F-K220).

It has been a long journey since I started my PhD in 2019. There have been many ups and downs along the way, including periods that were, perhaps, painful. Yet even at my lowest, there were always people beside me, accompanying and supporting me. I would not have been able to complete this thesis without your support.

I would like to start by thanking Mariette Berndsen – thank you for embarking on this amazing journey with me! Since my Honours year, you have been an amazing supervisor and mentor who not only taught me how to research, but also shed light on the many possibilities of who I can become – not just career-wise, but as someone who can trust myself to pursue what I want in life. Thank you so much for your support, encouragement, and, most importantly, your trust. Thank you for the faith you always had in me, even when I couldn't see it myself.

Thank you, Lydia Woodyatt, for always encouraging me to challenge myself and get the most out of my PhD experience. If it weren't for you, I wouldn't have achieved as much as I have, both within and beyond my research. You truly care about my personal growth, and you taught me to see the broader picture and look beyond my thesis. Without the adventures you led me to, my PhD journey would not have been this fulfilling and enjoyable.

Thank you, Emma Thomas, for always being caring—and especially for supporting me through the final stage of my PhD. It was an especially hard time for me, and you helped me refocus on my thesis and push through the toughest part. I am truly grateful for your support, and for the expertise and knowledge you contributed to my work. Your calm guidance kept me grounded.

Thank you, Michael Wenzel and Marika Tiggemann, for the insights you provided on my research proposal. Michael, I would also like to thank you and Lydia for creating such an amazing lab – the JEM Lab! Thank you, all my lab members, for your peer support and companionship – now that my thesis is done, I have no excuse not to join the Parkruns, right? (And drinks at the Tavern; let's stay healthy both physically and mentally.) I would particularly like to thank Anna Barron for providing valuable feedback on my conference presentations.

I would like to thank another mentor of mine, Delia Lin, who provided me with insightful guidance prior to and outside of my PhD – wisdom that proved invaluable in conquering this thesis.

To my family: thank you for supporting my decision to pursue a PhD. It is a privilege. Mum, Dad, and my sister – thank you for always encouraging me to chase my dreams. Tim, thank you for being an amazing stepdad. To my grandparents, who believed in the value of education – your beliefs have shaped how I view the world and myself, and I hold them close to my heart. And to Rambo, my family dog – thank you for keeping me company through the countless days and nights of writing. You mean so much to me, and I will remember you in my heart forever.

I would like to thank all my friends at and outside of Flinders University – especially Cindy, Doris, Eliana, Feng, Georgie, Henry, Jenn, Renyi, Siyu, Willa, and Yunhe. Thank you for always being there for me. I enjoyed every moment we spent together – study sessions, chats, walks, travels, drinks, hotpots. I once thought conference presentations were the highlights of my PhD, but that isn't completely true – the time spent with you is truly the highlight. Thank you for every moment we created together. I look forward to making even more fun and exciting memories with you.

Last but not least, thank you, Simon, for being an amazing partner. You supported me in so many ways—hanging out with me, playing Switch, crying together, watching Korean dramas, proofreading chapters… I look forward to the day you do a PhD so I can return the favour :)

List of Figures

List of Tables

A Note About the Format of the Thesis

This thesis has been prepared as a series of papers to be submitted for publication. Chapter 2-4 are manuscripts in preparation. Chapters 1 and 5 have been prepared in a traditional thesis format to provide context to the thesis as a whole. Given the format of this thesis, I have opted to use first-person singular pronouns ("I") in Chapters 1 and 5, and collective pronouns ("we", "our") to acknowledge the involvement of co-authors in the work to be submitted for publication (Chapter 2-4). In addition, to avoid repetition, I have created one single reference list that can be found at the back of the thesis.

CHAPTER 1. Introduction

During the 2013 Python Programming Conference (PyCon), a female technology advocate, Adria Richards, overheard two male developers behind her making jokes she found offensive and sexist. The men used technical terms like “dongles” and “forking,” which can carry sexual connotations when taken out of context (Ronson, 2016). Feeling uncomfortable, Richards took a photo of the men and tweeted about the incident, commenting on how such behaviour contributes to a hostile environment for women in technology: “Not cool. Jokes about forking repo’s in a sexual way and ‘big’ dongles right behind me. #pycon.” (Ronson, 2016, p. 108). Richards engaged in *online shaming*, a term defined here as the behaviour of publicly calling out someone or someone’s wrongdoing on social media (Barron et al., 2023; Billingham & Parr, 2020). Her action of condemning the joke had an immediate effect – shortly after she tweeted, the two male developers were asked to leave the conference by the staff. However, the power of shaming on social media was far beyond what Richards or many others anticipated. The situation eventually escalated, resulting in one of the male developers being fired from work (Starr, 2013).

In addition to highlighting the power of online shaming, the incident also exposed the divided views on the online shaming practice. Following Richards’ tweet, online discussion was divided into those who supported her actions and those who condemned her actions (Cutler, 2013; Fitts, 2017; greenrd, 2013; Selah, 2013). Some defended her right to call out what she perceived as being offensive, arguing that her action of standing up and confronting the male developers about sexism were necessary, justifiable, and even noble. Others criticised her for overreacting, disproportionately punishing the two men, and blaming her for the job loss of one of the men. The divided views fuelled the escalation of the incident and amplified the negative consequences of the online shaming. Adding to the controversy, in addition to the public condemnation of the male developers and the job loss for one of them,

Richards herself also became the target of online shaming and faced harassment, doxing (the public disclose of personal information), threats of violence, as well as her dismissal (Ronson, 2016; Franklin, 2013).

Of this example of online shaming, two key observations can be made: First, there were clearly groups emerging in relation to the shaming event, that is, the group who supported Richards and the group who condemned Richards. Second, the two groups of people seemed to share different goals as they contributed to this incident of online shaming. The group who supported Richards (including Richards herself) demonstrated a goal of raising awareness to address sexism in the tech industry (Gannes, 2013). The other group who condemned Richards wanted not only justice for the male developers who it perceived to have been unfairly shamed, but also punishment of Richards for her role in instigating action against him (Ronson, 2016). However, the emergence of groups and the presence of different goals these groups demonstrate have not yet been examined in the literature. The current thesis addresses such gaps in the existing online shaming literature by examining the research question: *Why do people engage in online shaming*? Central to this research question, I understand online shaming as a group behaviour that is influenced by group processes and the goal(s) pursued, individually and in combination.

The Importance of Understanding Why People Engage in Online Shaming

The debate about the appropriateness and use of online shaming observed from Richards' example can be consistently observed in other examples of online shaming (Saad, 2019), as well as from the broader media coverage and public discussions on online shaming (Muir et al., 2021). For example, a survey that examined American participants' views on "cancel culture," a phenomenon involving online shaming and social exclusion, revealed public debates around what online shaming was and why people engaged in online shaming (Vogels et al., 2021). The survey showed that some respondents viewed online shaming as a

potential means to hold the wrongdoer(s) accountable, while others viewed it as an unfair and undeserving punishment that sometimes involves censorship. I suggest that underlying these debates are different understandings of *why* people engage in online shaming.

Despite the growing public attention on online shaming and related concepts (Vogels, 2022), the literature examining this phenomenon is highly interdisciplinary in nature and lacks empirical research on the motives which might drive people's engagement in online shaming (Muir et al., 2023), with some psychological research only starting to emerge recently (e.g., Barron et al., 2023; Hou et al., 2017; Muir et al., 2023). However, in contrast to the relatively underexplored reasons behind why people engage in online shaming, there is growing attention from both academia and technology industry on mitigating this practice (Bodaghi et al., 2023; Li et al., 2024; Record & Miller, 2022). For example, some researchers examined methods to classify and detect online shaming comments (Surani & Mangrulkar, 2021), while others developed tools to reduce such comments on social media (Basak et al., 2019).

Given these recent trends around reducing online shaming practice, I suggest that there is an urgency to examine the motives of people who engage in online shaming. Exploring this question not only addresses the growing interest and ongoing debates surrounding the practice of online shaming, but also holds practical significance. Recently, for example, the Singaporean government has considered introducing legislation to ban cancel culture and online shaming (Jalal, 2023). Therefore, I suggest that understanding the reasons behind people's engagement in online shaming can inform various social actors – such as individuals, groups, social media platforms, and authorities – help them respond appropriately to the practice of online shaming and to those who engage in it. Specifically, through conceptualising online shaming as a group behaviour, I propose that, although online

shaming might involve a motive to harm others, it can also be driven by collective goals that are shared among a group of people.

Conceptualising Online Shaming as Group Behaviour

Despite growing attention on the online shaming phenomenon, there is a lack of academic consensus on its conceptualisation. I offered a working definition of online shaming in the opening paragraph above, but others have considered online behaviour of similar kinds through very different conceptual lenses. For example, consistent with the line of research aimed at reducing online shaming, some researchers have conceptualised online shaming as a form of online aggression or toxicity, emphasising its abusiveness and an individual's intent to harm others (e.g., Amit-Aharon et al., 2023; Behera et al., 2022; Ge, 2020; Hou et al., 2017). In contrast, other researchers suggested that online shaming can entail positive motives, such as seeking justice and raising awareness of a social issue (as shown from Richards' example) (Blitvich, 2022; Kitchin et al., 2020). However, in general, there is still a lack of empirical research examining the psychological motives behind people's engagement in online shaming.

Perhaps most importantly for this thesis, research has focused primarily on understanding online shaming from an individual's perspective, which includes examining psychological and situational factors such as personality traits, personal beliefs and judgements on morality and justice, as well as social media usage (e.g., Ge, 2020; Hou et al., 2017; Muir et al., 2023; Pundak, 2021). I draw on a social psychological perspective to argue that understanding online shaming solely through individual behaviour may be insufficient. Recall the Richards' example, which I suggested that it involved two groups of people demonstrating different collective goals. It not only shows that online shaming can be guided by more positive goals rather than just causing harm, but also highlights that motives and goals can be shared among a group of people. Therefore, consistent with this rationale, the

current thesis adopts a social psychological lens and conceptualises online shaming as a group behaviour. Instead of viewing people who engage in online shaming as isolated individuals with hurtful intent, I draw from the social identity approach (Tajfel & Turner, 1979; Turner, 1985) and conceptualise the participants of online shaming as part of a social group in an online environment within a broader social context.

Online Shaming Involves Group Processes

The current thesis adopts the social identity approach to examine why people engage in online shaming. Comprising of Social Identity Theory (Tajfel & Turner, 1979; Turner, 1982/2010) and Self-Categorisation Theory (Turner et al., 1987), the *social identity approach* suggests that people can view and define themselves according to the group they belong to. When a group identity is important and salient to individuals, it can influence how they perceive and navigate their social environment. Specifically, it can lead individuals to internalise the norms and the values of the ingroup, which in turn, influences their feelings, perceptions, attitudes, and even behaviours in ways that are consistent with the group norms and expectations. When their group identity is under threat, those who are committed to the group are particularly motivated to affirm the ingroup, which can result in the group behaviours that challenge the source of threat, improve the status and image of the ingroup, as well as display their group affiliation (Doosje et al., 1999; Ellemers et al., 2002). Aligning with these arguments put forward by the social identity approach, I argue below that online shaming can be a way of addressing an identity threat.

Online shaming is usually triggered by a norm transgression that involves a value violation (Barron et al., 2023; Blitvich, 2022; Haugh, 2022). These underlying values often prescribe what people should or should not do in a given context (Cialdini et al., 1991). In other words, they provide a reference on what people believe as morally right or wrong, which can be determined and influenced by one's group identity (Ellemers & van den Bos,

2012; Spears, 2021). In the example of Richards, the male developers' behaviour of making jokes with sexual connotation may be perceived as having violated the norms relating to the underlying values of professionalism, respect, and gender equality in a work environment (Cutler, 2013). Moreover, some might perceive such violation as worthy of being called out, as they hold those values of particular importance. In the same example, Richards' identity of being a woman who works in a male-dominant tech industry was likely salient (Milstein, 2013), especially at a conference stressed on an inclusive environment (Tollervey, 2013). Therefore, exposing the norm violation as well as shaming the norm violators, could have been perceived by the ingroup as necessary means to restore the violated value and affirm one's ingroup (Brady et al., 2020; Ellemers et al., 2002).

Online shaming is likely to involve intergroup interactions and conflicts between different group values. The exposure of norm violation can invite fellow ingroup members who share the same view to join the condemnation of the value-violating behaviour (Haugh, 2022). However, it is also likely to attract condemnation from outgroup members who disagree on interpreting the exposed behaviour as a norm violation therefore oppose the online shaming behaviour (Adkins, 2019). Such intergroup conflicts are explained by the fact that norms are often contested and as a consequence, what constitutes a norm violation can be determined by one's group membership (Spears, 2021). Consider the group of people who shamed Richards for her call-out of the male developers. Many of them did not perceive the male developers' jokes as inappropriate, therefore the the developers' behaviour was not a norm violation, let alone deserving severe punishment such as viral shaming and the job loss.

Furthermore, online shaming is a dynamic process that evolves over time. Specifically, the action of one group can shape the social identity and the context for another group (Reicher, 1984), leading to dynamic interplay that can further escalate or deescalate shaming. For example, the dismissal of Richards might have triggered further outrage of the

ingroup, which led to further derogation of the outgroup. In other cases of online shaming, people often create new hashtags to express their support for someone as well as condemnation towards others, such as the hashtag #IStandWith followed by the person or group that they supported (e.g., Saad, 2019). Consistent with the research that examined how self-labelling with hashtags could (re)define a social identity and lead to group-based behaviours such as further activism (Barron & Bollen, 2022; Foster et al., 2020), I suggest that group identity and intergroup relationship can be redefined and transformed as online shaming progresses.

Online Shaming Involves Pursuit of Goals

Implied in the discussion above is the conceptualisation of online shaming as a group behaviour, entailing shared goals that people pursue as a group. When a group has its own agenda, online shaming can be used strategically to achieve a collective goal. For example, in cases where justice cannot be sought through formal channels, online shaming can be used as an alternative way to achieve justice by publicly sanctioning the perceived wrongdoer (e.g., Leopold et al., 2021; Jane, 2017). Although it is debatable whether the jokes from the male developers should be interpreted as inappropriate, such example of online shaming also showed a shared goal of improving work culture and raising awareness on the challenges (such as sexism) faced by females working in the tech industry (Cutler, 2013). This suggests that there could be other goals, besides causing harm, in at least some online shaming incidents.

However, there is still a lack of research that examines the overarching psychological goals that account for not just individual incidents of online shaming, but also online shaming behaviour in general. Some psychological research on online shaming has shown that, in addition to merely causing harm, online shaming could also be driven by motives to benefit the society and/or seek justice (e.g., Hou et al., 2017; Skoric et al., 2010). These studies often

take a more individual approach and view those who engage in shaming as isolated individuals, instead of conceptualising online shaming as a group behaviour that reflects collective goals. Furthermore, without integration in the broader, multidisciplinary literature, it remains unknown whether there might be other motives exist in online shaming.

To address these gaps, the current thesis examines the multidisciplinary literature and maps out the overarching psychological goals behind the motives driving people's online shaming behaviour. It is important to note that I commence with an inductive approach to identify the goals that are relevant to online shaming. Specifically, my thesis was not guided by a goal/motivational theory such as goal-setting theory (Locke & Latham, 1990) or self-determination theory (Ryan et al., 2000). Instead, I adopt a bottom-up approach to explore and classify in the multidisciplinary literature the reasons why people are understood to engage in online shaming. As I explain below, this approach is appropriate given the nature of the multidisciplinary literature and allows for a more integrated understanding of the complex motivations underlying online shaming behaviour.

The Current Thesis

In recent years, the role of online shaming has been much debated in the public, the media, and academic communities (Muir et al., 2021; Vogels et al., 2021). In contrast to the prevalence of online shaming in everyday social interactions and popular interest in its practice, the literature in psychology on online shaming remains limited. Drawing on the insights from past online shaming events and their related debates, my PhD research adopts a social identity approach and conceptualises online shaming as a group behaviour. It involves both group processes and the pursuit of goals, which explains why people engage in online shaming, individually and in combination. Specifically, the current thesis is comprised of four studies: a scoping review that offers an integration of the multidisciplinary literature (Chapter 2), a study of a real-life shaming event on Twitter (Chapter 3), and two experimental studies

(Chapter 4). A brief overview of the thesis outlining the objectives of each study is presented in the text below.

Overview of Chapters

Chapter 2 presents a scoping review (Study 1) of the multidisciplinary literature that examines why people engage in online shaming. The purpose of the scoping review is to systematically map out the potential motives of online shaming that are identified in the literature. I am also interested in examining how online shaming has been conceptualised and studied in this topic. The review includes a thematic synthesis that offers a way of integrating the multidisciplinary literature. With an inductive approach, four overarching psychological goals that could drive online shaming are identified from the literature. The four goals are: 1) punishing the perceived wrongdoer, 2) deterring the perceived wrongdoing, 3) seeking social acknowledgement, and 4) creating change. The scoping review informs the conceptualisation of online shaming that is developed and empirically tested in the subsequent chapters.

In Chapter 3, I investigate whether the goals identified from Chapter 2 are presented in a real-life online shaming event (Study 2). With analysing a two-phased shaming event occurred on Twitter, the study provides an opportunity to examine how the goals emerge and develop as online shaming progresses over time. The psychological goals identified from the previous chapter (Chapter 2) were generally supported. It was also found that the progression of online shaming was influenced by group formation and intergroup dynamics, supporting our claim that online shaming should be understood as a group behaviour that can be driven by multiple goals.

Chapter 4 consists of two experimental studies that are designed to examine whether group dynamics influence people's online shaming engagement. I investigate whether people can be mobilised by a leader through a shared social identity to engage in online shaming, that is, becoming engaged followers to punish the perceived wrongdoer. I used two

paradigms where the aim of online shaming was to elicit certain types of prosocial behaviour of others. In Study 3, the goal was set to reduce online hostility, with participants either being introduced to a mobilising leader (i.e., who belongs to an ingroup, presents a punitive norm and a noble goal to support the punishment), or a non-mobilising leader (i.e., who belongs to an outgroup and lacks any explicit message about the punitive norm or the noble goal). Study 4 was conducted during the emergence of COVID-19, with a goal set to encourage the compliance to the COVID-19 guidelines. To further investigate the leader's mobilisation, the leader's social identity (whether they belong to the ingroup or the outgroup) and the level of mobilisation (whether a noble goal and a shaming norm is present) were separately manipulated in Study 4.

Chapter 5 features a general discussion of the thesis. I present a summary and synthesis of the key findings of the studies, an integrated account that explains why people engage in online shaming, followed by the discussion of the strengths and limitations of the research. I then discuss how future research can further address the group processes and goals that drive people's online shaming engagement, as well as future directions to address the online shaming phenomenon. The chapter ends with a discussion on the implications of the current research.

Authorship Statement

Chapter 2 is based on a co-authored manuscript preparing for future publication:

Zhao, K., Berndsen, M., Woodyatt, L., Thomas, E. (2025). Why do people engage in online shaming? A scoping review of multidisciplinary literature. [Unpublished manuscript]. Flinders University

The candidate was the primary author of the work. Specifically, the candidate was responsible for:

- Research design,
- Data collection and analysis, and
- Manuscript writing and editing.

Co-authors provided supervision, critical feedback on study design, data collection and analysis, and results interpretation during the writing process. Specifically, the second author assisted as a second coder screening the literature. This work has been included in the thesis with the permission of the co-authors.

CHAPTER 2. Why Do People Engage in Online Shaming? A Scoping Review of Multidisciplinary Literature

Abstract

Online shaming, the behaviour of calling out someone's perceived wrongdoing on the internet, has attracted widespread discussion and considerable media attention because of the negative consequences it can cause. While online shaming has been discussed in many disciplines, such as sociology, law, media and communications, and computer science, there is no consensus on what drives people to engage in online shaming. To explore the current research on the psychological motives of online shaming, we systematically searched the multidisciplinary literature using a variety of databases, including Web of Science, ProQuest, Scopus, PsycINFO, Google Scholar, to gather relevant research records (N = 94). Through systematically mapping out the potential motives of online shaming, we identified four psychological goals that can drive people's online shaming engagement: punishing the perceived wrongdoer, deterring the perceived wrongdoing, seeking social acknowledgement, and creating change. Our review suggests that, rather than viewing online shaming as merely an individual act of hurting others, we need to understand and examine online shaming as 1) a group behaviour in an intergroup context, and 2) a behaviour that can sometimes be driven by goals in addition to harm, such as creating social change, which may be perceived as noble social goals – at least by some.

Introduction

Public shaming, the practice of publicly humiliating and punishing deviants, has deep historical roots and been used across different societies, such as the tarring and feathering used in medieval Europe and American colonies (Levy, 2011), and denunciation rallies occurred during the Chinese Cultural Revolution (Lu, 2004). More recently, with the growth of social media (Matei, 2019), public shaming can now be done almost effortlessly on the internet by calling out someone or someone's wrongdoing with responses such as liking, sharing, or commenting on a social media post (Barron et al., 2023; Basak et al., 2019; Billingham & Parr, 2020). Indeed, the prevalence of online shaming can be demonstrated from a survey conducted in 2016, where it was found that 60.3% of the participants have engaged in online shaming at least once over the last two months, and 25.1% of the participants reported of being both a victim and a perpetrator of online shaming for the same period of time (Packiarajah, n.d.).

In contrast to the effortlessness of engaging in shaming online, the consequences on those who experience shaming remain substantial. In instances of massive online shaming, those who experienced shaming faced severe consequences, such as being fired, dropping out of university, or even taking their own life (Krim, 2005; Kubovich, 2015; O'Neill, 2017). Online shaming of smaller scale can still cause long-lasting consequences for those who are shamed, such as withdrawing from using social media, self-surveillance and self-censorship (Huffman, 2016; Laywine, 2021; Marwick, 2021). For its ubiquity and negative consequences, online shaming has attracted both media and researchers' attention (Muir et al., 2021). However, there is still little consensus on the conceptualisation of online shaming as well as why people engage in online shaming, and research on online shaming remains interdisciplinary.

Early research on online shaming predominantly emphasised its punishing nature (e.g., Cheung, 2014; Klonick, 2016; Laidlaw, 2017; Wehmhoener, 2010). Such an emphasis on the abusiveness and hurtfulness of online shaming is reflected from the conceptualisation of online shaming being a type of cyberviolence or aggressive behaviour (e.g., Ge, 2020; Huffman, 2016). However, recent studies have not only continued recognising the punishing nature of shaming but also begun to explore the interplay of shaming with other psychological factors that might explain people's intention to punish/shame. These factors include reputational concerns (Johnen et al., 2018), one's justice concerns and beliefs (Chang & Poon, 2017; Hou et al., 2017; Rost et al., 2016), psychopathic tendencies (Muir et al., 2023), as well as moral emotions such as schadenfreude (pleasure about someone's misfortune), anger, and outrage (Barron et al., 2023; Blitvich, 2022).

While research examining online shaming as an act of violence or aggression continued to evolve (Amit-Aharon et al., 2023; Ge, 2020; Šincek, 2021), some researchers also began to acknowledge that online shaming can sometimes be driven by group-based motivations and/or concerns for the society at large (e.g., Brady et al., 2020; Gruber et al., 2020; Haugh, 2022; Leopold et al., 2021). These different lines of research applied different understandings to online shaming, and implied a changing view among researchers on why people would engage in online shaming. In the current scoping review, we aim to systematically map out the potential motives of online shaming that are represented by various multidisciplinary literature, as well as to examine how researchers' understanding of online shaming has changed over time.

Scoping review is often used to provide an overview of a body of literature that covers emerging evidence (Munn et al., 2018). We suggest that a scoping review is more appropriate than a systematic review or a qualitative evidence synthesis (Grant & Booth, 2009), because of the interdisciplinary nature of online shaming literature, the presence of

both quantitative and qualitative studies, as well as the aims to identify motives of online shaming. The steps taken to conduct the scoping review are described below. First, we identified research records that were related to online shaming based on our exclusion criteria, and extracted data that was related to the research question (i.e., why people engaged in online shaming). This was then followed by a thematic analysis on the literature, mainly based on the data extracted from the included research records. Specifically, the thematic analysis was guided by examining recurring patterns (themes) within the literature to provide insights of people's goals when engaging in online shaming (Braun & Clarke, 2006). We also examined how the presence of these themes emerged and changed in the literature.

Method

We systematically searched the literature across different disciplines and applied screening techniques to identify the relevant research. An initial search was conducted through four databases, Web of Science, ProQuest, Scopus, and PsycINFO. Pilot searches with key terms, such as *internet shaming* and *online shaming*, were conducted by the primary researcher with the help of a librarian. We then undertook an iterative approach by browsing the articles that appeared in the pilot search results and adding more keywords to the list of the search terms. The final key terms included: *internet shaming*, *viral outrage*, *social media firework*, *digital vigilantism*/*vigilante*, *online humiliation*, and the variations of these terms, such as *online outrage*, *digital shaming*, *cyber vigilantism*. Since there was no consensus on the definition of online shaming, related concepts such as *online disgust*, *online contempt*, *doxing* (e.g., Brady et al., 2021; Douglas, 2016; Moore, 2016) were also included in the

initial search.[1] We also searched the grey literature that was not published in the commercial journals (e.g., theses or dissertations and conference papers; Paez, 2017).

Through screening the title, abstract (or full text if the abstract was not available or not sufficiently informative) and keywords, a record was included when it met the following criteria:

1) published or was made available between January 1, 2000, and July 30, 2019. As there were more publications of online shaming made available since the initial search, two follow-up searches were conducted on July 15−16, 2020 and on March 4, 2024, respectively. The first follow-up search was conducted via ProQuest, Scopus, Google Scholar, and by checking the reference lists of relevant journal articles. The second follow-up search was conducted via ProQuest, Scopus, and by checking the reference lists of relevant journal articles;
2) the full text of the research was available in English;
3) the aim or the research question was relevant to online shaming targeting individual(s), and online shaming was the focus of the research;
4) the internet was the key medium of shaming; and
5) online shaming did not merely happen within close interpersonal relationships (for instance, parent-shaming and partner-shaming do not meet this inclusion criterion).

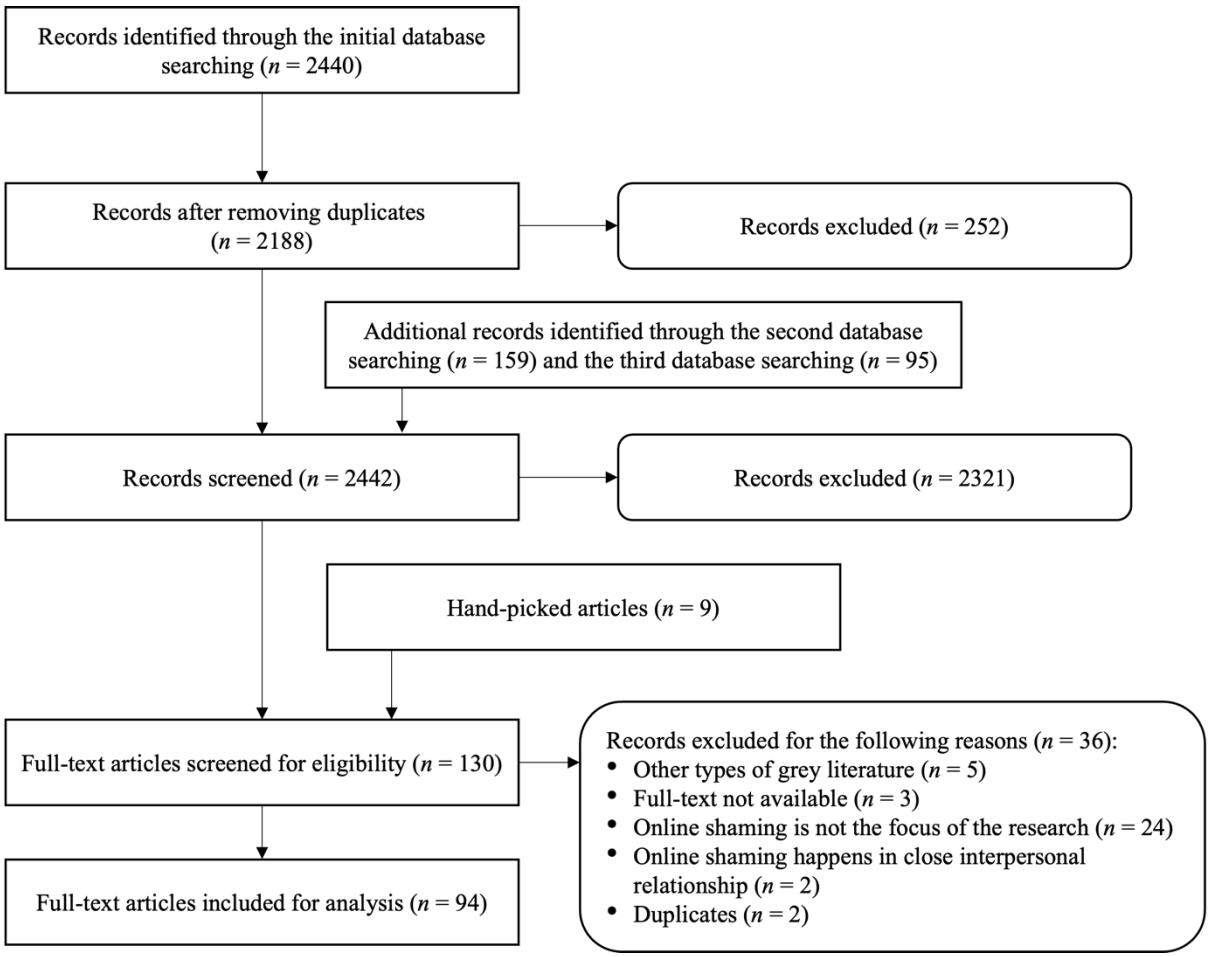
A total of 94 records were included and reviewed (see Figure 1 for the flowchart of the screening process). As shown in Table A1 in the Appendix A, the majority of the included

[1] See Appendix A for a complete list of the search terms for the initial and the follow-up searches. After the initial search, we had a clearer definition on online shaming, thus reduced the search terms in the follow-up searches.

records are peer-reviewed journal articles (n = 78), along with graduate theses or dissertations (n = 8), commentaries or letters published in academic journals (n = 4), book chapters (n = 2), a conference paper (n = 1), and a non-peer-reviewed journal article (n = 1). Following Peters et al.'s (2020) recommendations for conducting a scoping review, the primary researcher extracted the following data from each record: the discipline of the research, definition of online shaming, focus of the study or research question(s), method(s), and key finding(s) or argument(s). The primary and the secondary researchers then coded the included records independently from each other into preliminary categories based on the extracted data, particularly of how online shaming was understood by researchers and the identified reasons why people might engage in online shaming behaviour. Based on the preliminary categories, initial classifications were formed through discussion among three researchers and agreement was established. These initial classifications served as a basis for our subsequent thematic analysis, that is, the goals that might drive people's engagement in online shaming. Before outlining the results of the thematic analysis, we first present the descriptive findings below, including the number of articles published in each discipline and the research methods used by the researchers to study online shaming. This would then be followed by the researchers' conceptualisation of online shaming.

Figure 1

Flowchart of Screening Process



Findings

Descriptive Findings

Among the included records ($N = 94$)[2], most of them came from the disciplines of media and communication ($n = 32$), psychology ($n = 25$), sociology ($n = 12$), criminology ($n = 10$), law ($n = 9$), philosophy ($n = 8$), and computer science ($n = 8$). The remaining records ($n = 13$) came from other disciplines: feminist studies, political science, management, sport, business, and education. Both qualitative and quantitative methods (and sometimes mix-methods) were used to study online shaming ($n = 53$). The quantitative research in the included articles primarily employed self-reported surveys and experiments to examine people's experiences of online shaming, factors influencing people's perception and attitude towards online shaming, and factors influencing people's engagement in online shaming. The qualitative research primarily applied thematic analysis, content analysis, and critical discourse analysis to analyse qualitative surveys, interviews, and archival data (such as comments or replies that people made online, the content of a certain online forum or discussion board, and news that reported online shaming). Most of the qualitative studies involved studying either a single or several case(s) of online shaming. And lastly, the mixed-method research primarily used either a combination of statistical analysis and content analysis to analyse archival materials or involved separate qualitative and quantitative studies.

[2] As shown in Table A1, some of the included records are cross-disciplinary research, thus, they are classified into multiple disciplines.

Definition of Online Shaming

An examination of the extracted definitions revealed that researchers defined online shaming in various ways (see Table 1 for example definitions), including cyberviolence, punishment/vigilantism (or "human flesh search" in the Chinese context, e.g., Bu, 2013; Gao, 2013), whistleblowing behaviour, expression of outrage, and strategy for activism. These different conceptualisations of online shaming seem to have emphasised the different aspects of online shaming (Table 1). More specifically, the involvement of hurting and abusing is captured in the definitions of online shaming as a cyberviolence, as well as a punishment and/or an online vigilantism (e.g., Dilmaç, 2014; Klonick, 2016). These definitions also emphasised the public nature of online shaming (e.g., Cheung, 2014; Chia, 2019; Huffman, 2016), with the conceptualisation of online shaming as a punishment (vigilantism) having a particular focus on norm enforcement and achieving social justice (Klonick, 2016). Similarly, Skoric et al. (2010) interviewed organisers behind shaming blogs and YouTube channels, and described them as whistleblowers who called out their peers for violating social norms. While this definition effectively describes how illegitimate or immoral behaviour can be exposed by those who *initiated* the shaming, the motives might differ among individuals who join the shaming after it has already gained widespread attention. This is because the shaming behaviour at this point no longer serves the function of disclosing new information as has been the case in the whistleblowing behaviour. Instead, it might now be better described as an expression of moral outrage (or an online firestorm), where someone is condemned by a large number of individuals for an offensive remark or behaviour. Finally, some researchers understood shaming as a strategy for online activism, highlighting its potential to bring together a group of people towards a common objective and to foster social changes.

Table 1

Unique Conceptualisations of Online Shaming and Examples Proposed in the Reviewed Articles

Unique conceptualisation of online shaming	Example quote
A type of cyberviolence	"Therefore, as Cheung (2014) has described, online shaming is an act of violence occurring when individuals perceived to have transgressed social or moral boundaries are persecuted by the anonymous crowd on the Internet for the purpose of public humiliation" (Huffman, 2016, p. 12)
An informal, third-party punishment (including online vigilantism)	"Accordingly, online shaming is (1) an over-determined punishment with indeterminate social meaning; (2) not a calibrated or measured form of punishment; and (3) of little or questionable accuracy in who and what it punishes" (Klonick, 2016, pp. 1029–1030) "Just like cyber bullying or harassment, it often involves repeated verbal aggression over time, but it has another key element: shaming also involves the attempt to enforce either a real, or perceived, violation of a social norm." (Klonick, 2016, p. 1034)
A whistleblowing behaviour	"The act of contributing materials to online shaming websites can be regarded as an act of whistle blowing on individuals who behave in ways contrary to social norms" (Skoric et al., 2010, p. 185)
An expression of moral outrage	"Expressing moral outrage—a combination of anger and disgust at the violation of a moral standard (Salerno & Peter-Hagene, 2013)—communicates to others that the violator is reprehensible (Crockett, 2017). … This may be one motivation behind viral outrage, where an individual's offensive remark inspires online condemnation from thousands, sometimes millions (Ronson, 2015)" (Sawaoka & Monin, 2020, p. 499)
A form of collective action (or a tactic for activism)	"Shaming as feminist discursive activism" (Abraham, 2014, p. 166)

To further understand which aspects of online shaming were emphasised predominantly in the literature, we plotted a word cloud of the most frequent words or phrases appeared in the definitions extracted from the included records. With words pre-processed, a word cloud showing the most frequent 50 words/phrases was generated using *R* (see Figure 2). Consistent with the conceptualisation of online shaming as a punishment/vigilantism, the public nature of online shaming was emphasised the most (as demonstrated by the keywords "public" and "publicly"). Words such as "vigilantism", "punish", "punishment", "human flesh search" and "justice" also appeared frequently in the definitions. The moral aspects of online shaming, such as moral outrage (e.g., "moral", "outrage"), involvement of perceived transgression and social norm violation/enforcement (e.g., "social norm", "violated", "transgression"), information searching and dissemination (e.g., "information" and "human flesh search"), the medium of online shaming (e.g., "internet" and "social media"), as well as the social actors involved in shaming (e.g., "people", "individuals", "target", "others"), were also prevalent themes in the definitions. The hurting and abusing aspect of online shaming can also be reflected from a range of words, such as "shame", "harassment", "humiliate", "slut-shaming", and "doxing". Lastly, as shown from the words "type" and "form", some researchers focused on a specific type or case of online shaming (e.g., slut-shaming, Papp et al., 2017; Papp et al., 2015), or suggested that online shaming be encapsulated by broader concepts such as expression of moral outrage and vigilantism (e.g., Brady et al., 2020; Dunsby & Howes, 2019; Gao, 2013). Drawing on the overlaps of the various conceptualisations of online shaming as well as the overarching themes appeared in the definitions, we expand Braithwaite's (1989) definition of shaming and define online shaming as a behaviour by which *individuals communicate their disapproval of a perceived wrongdoer's violation of a norm or value via an online public platform*.

Figure 2

Word Cloud for the Extracted Definitions

Note. The most frequent words "online" and "shaming" were excluded to reveal what "online shaming" represents. The sizes of the words or phrases represent their frequencies, with more frequent words plotted larger. Words or phrases with the similar frequencies were grouped into the same colour.

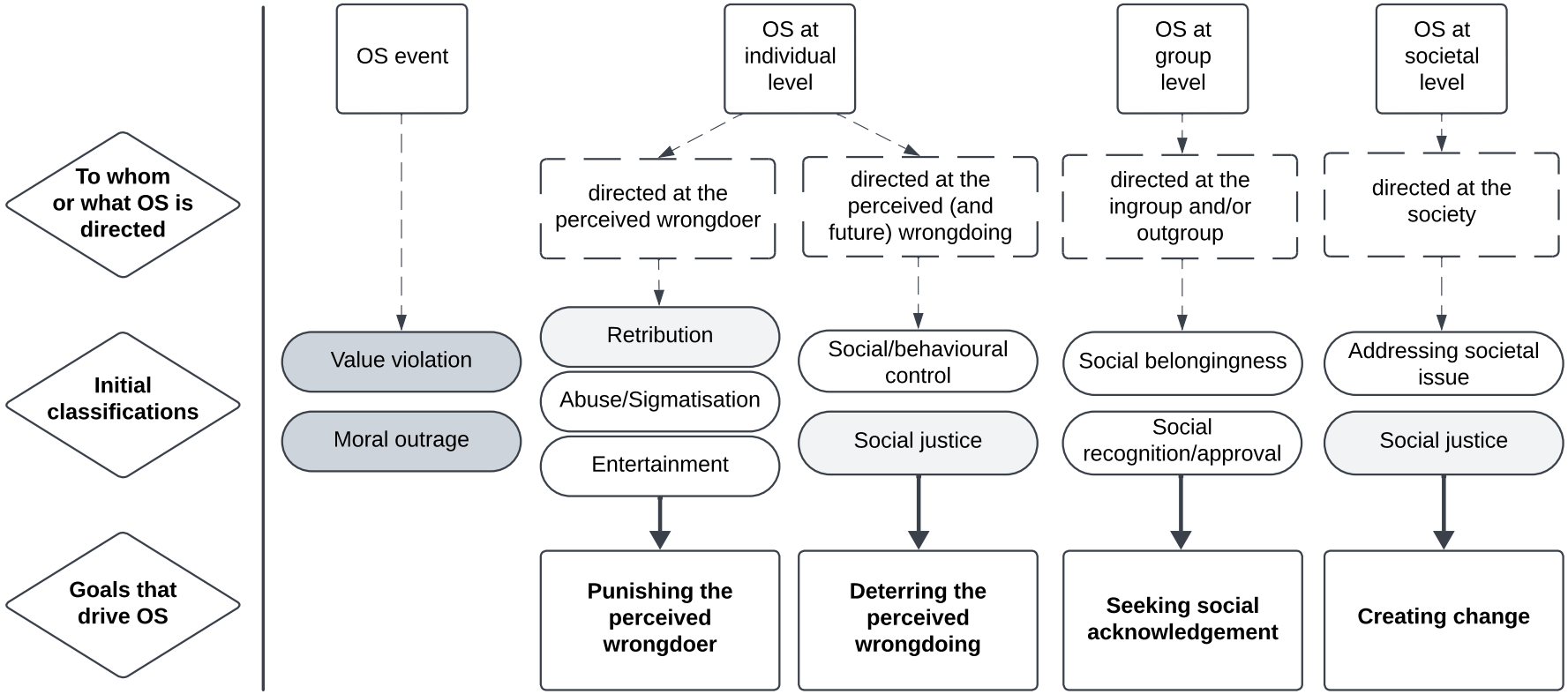
Thematic Analysis

Based on the preliminary categories identified in the extracted data from the included records, we formed some initial classifications to identify reasons why people engaged in online shaming behaviour. As shown in Figure 3, most initial classifications can be mapped onto four goals. They are the goal of 1) punishing the perceived wrongdoer, 2) deterring perceived wrongdoing, 3) seeking social acknowledgement, and 4) creating change. We took a deductive approach to the thematic analysis, guided by social psychological theories that informed our understanding of human response to injustice, including Braithwaite's (1989) stigmatising and reintegrative shaming, just-desert theory and deterrence theory (Carlsmith et

al., 2002; Carlsmith, 2006), as well as the social identity approach (Tajfel, 1978; Tajfel & Turner, 1979). It is worth noting that, while the four goals are presented individually in Figure 3, these goals often intersect with one another. Below, we will introduce each of the goals in detail and describe how they are presented in the literature.

Figure 3

Goals of Online Shaming Based on the Initial Classifications



Note. OS = online shaming. The current figure presents the initial classifications that were related to the goals why people engage in online shaming. The initial classifications (in the pill-shaped elements) with different shades (i.e., *Value violation*, *Moral outrage*, and *Social justice*) can be mapped onto more than one goal.

Why do People engage in online shaming?

Punishing the Perceived Wrongdoer. We identified punishing the perceived wrongdoer as the first goal of online shaming. This goal involves actively seeking retribution against a perceived wrongdoer for their past norm violation. Particularly, online shaming out of this goal can be understood as a third-party punishment since people who engage in shaming are usually not directly harmed by the violation, nor seemed to receive any direct benefit from the shaming behaviour. The goal of punishing can be reflected from the widely acknowledged harm and abusiveness associated with online shaming across disciplines (e.g., Billingham & Parr, 2020; Cheung, 2014; Chia, 2019; Frye, 2021; Ge, 2020; Haugh, 2022). However, there are only a few studies empirically examined people's intent to inflict suffering on those who have violated social norms (e.g., Barron et al., 2023; Ge, 2020; Muir et al., 2023).

We suggest that the motives to seek retribution can be explained by just-desert theory, which refers to people believing that wrongdoers deserve punishment in proportion to the moral offensiveness of their action and for the past harm they have caused (Carlsmith et al., 2002; Carlsmith & Darley, 2008; Darley et al., 2000). It was found that online shaming can be motivated by people's justice concerns following a norm violation, via an increased attribution that the perceived wrongdoer deserves to be called out publicly and the associated negative consequences, as well as via an increased feeling of pleasure from the wrongdoer's suffering (i.e., schadenfreude) (Barron et al., 2023). This study provides evidence that online shaming can indeed be motivated by seeking retribution, which supports punishing the perceived wrongdoer as a goal that drive people's online shaming behaviour.

Online shaming might result in more severe consequences than the original offences seem to warrant (e.g., Billingham & Parr, 2020; Frye, 2021; Haugh, 2022), and in some cases being contradictory to the proportionality principle of retributive justice (whereby the

punishment fits the offence). Research on public shaming suggested that when it occurs within communities where there are pathways to repair, shaming can have the potential of educating wrongdoers and *reintegrating* the offender back to the community (Braithwaite, 1989). However, online shaming incidents, especially for large scale events, are usually disintegrative. Online shaming tends to involve exposing the perceived wrongdoer as a flawed *person*, attacking their dignity and excluding the perceived wrongdoer from one's group (Laidlaw, 2017; Wall & William, 2007). Therefore, different to certain public shaming that occurs within communities, large-scale online shaming events deprive the perceived wrongdoer's opportunity to be reintegrated to the community (Frye, 2021; Haugh, 2022; Wall & William, 2007).

A focus of research in the area has been to examine the underlying reasons why individuals persist in supporting or engaging in online shaming, even when such actions become disproportionate to the original norm violation. One reason could be that people who engaged in online shaming were unaware or downplay how hurtful online shaming might be, due to the situational circumstances or factors that could influence people's perception (Ge, 2020; Puryear & Vandello, 2019). In cases with severe consequences, the inflicted harm of online shaming still be overlooked by people who engaged in shaming as well as the public who observed shaming. For example, in a case where online shaming led a victim to take their own life, the Twitter responses discussing this case tended to portray the victim as inherently vulnerable and unable to protect themself (Thompson & Cover, 2022), rather than to emphasise the inflicted harm of shaming.

The online environment, especially social media, might have played a crucial role in contributing to the lack of awareness of the harm caused by online shaming, and the acceptability of the disproportionality of the punishment. It was found that social media usage led to the reduction in users' moral sensitivity, because users were less aware of the moral

intensity of hostile comments, in which the perceived potential harm was understood as a key dimension of moral intensity (Ge, 2020). This process gave rise to users' engagement in online shaming. Similarly, online shaming was found to be predicted by moral disengagement (Muir et al., 2023), the process where one distances themselves from their ethical standards due to perceived extenuating circumstances (Bandura, 1999). Together, this line of research examining online shaming suggested that people participating on online shaming are not always motivated by, or even aware of, the severity of punishment that can eventuate.

Researchers have also proposed that certain personality traits could explain, to some extent, why people choose to punish or make others suffer via online shaming. It was found that people's engagement in online shaming could be explained by certain personality factors, such as the dark triad including machiavellianism (the tendency of being exploitative for their self-interest), narcissism (the tendency to process a sense of entitled self-importance), and psychopathy (the tendency of being impulsive, selfish, and non-empathetic) (Muir et al., 2023). Specifically, psychopathy was found to be the strongest, unique predictor of participants' shaming intentions among the dark triad personalities, while empathy was found to negatively predict online shaming intentions. Along with other predictors such as social media usage and moral disengagement, these predictors, including the personality traits, were found to account for 39% of the variance in online shaming intentions, and 20% of the variance in people's perceived deservingness of online shaming. These findings suggest that there are certain individual differences that could explain the punitive nature of online shaming.

Taken together, these explanations (situational and personality traits) provided evidence that punishment is a goal for some in online shaming, and that there may be psychological mechanisms that enable people to participate in punishment despite the fact that it can inflict disproportionate harm to the perceived wrongdoer.

Deterring the Perceived Wrongdoing. We identified deterring the perceived wrongdoing as another goal that could explain why people engage in online shaming. Muir et al. (2023) found that, in addition to viewing online shaming as punishing someone, some people who engaged in shaming also considered their behaviour a deterrence to others' action, and believed that shaming could ensure accountability. The goal focuses on deterring both the wrongdoer and others from committing similar transgressions in the future. Such a goal can be supported by the deterrence theory of punishment, which justifies punishment by minimising the possibility of future offenses by increasing the cost of violating norms (Carlsmith, 2006). Therefore, different from the first identified goal of online shaming that emphasises on giving what the wrongdoer deserved (i.e., retribution for a past harm), the goal of deterring the perceived wrongdoing involves a forward-thinking and focuses on future compliance to a norm.

Online shaming is often criticised for being abusive and stigmatising, particularly in its goal to punish the perceived wrongdoer. However, researchers also recognised its potential to achieve positive social outcomes. Notably, across disciplinary boundaries, online shaming is widely regarded as a way to reinforce societal norms and deter similar violations in the future (Barron et al., 2023; Brady et al., 2021; Klonick, 2016; Mielczarek, 2018; Muir et al., 2023; Wehmhoener, 2010). For instance, a study on young people's perceptions of online shaming found that it was seen as distinct from bullying due to its perceived benefits, such as discouraging undesirable behaviours (de Vries, 2015). Consistently, online shaming has been framed as a form of social control aimed for deterring deviant behaviours, maintaining social order, and at times, contributing to the society and justice (e.g., Blitvich, 2022; Chia, 2019; Hou et al., 2017; Laidlaw; 2017; Skoric et al., 2010).

While the goal of deterring the perceived wrongdoing and the goal of punishing the perceived wrongdoer address different aspects of online shaming, they are linked through a

shared understanding of online shaming as a form of punishment. The relationship between these two goals can be supported by recent research that examined non-maleficence principle, that one should avoid causing harm to others intentionally (Pundak et al., 2021). On one hand, adhering to such a principle could be negatively associated with one's participation in shaming when shaming is viewed as being harmful. On the other hand, keeping to this principle might also encourage people to engage in shaming, particularly if shaming is viewed as a means to deter similar future wrongdoing. The research findings by Pundak et al. showed that people who had a higher adherence to non-maleficence principle were more likely to engage in online shaming, but only when the identifiability of the perceived wrongdoer was low, that is, when the inflicted harm was expected to be milder. Therefore, when it comes to online shaming, people are concerned with both the consequences in relation to the past harm done (retribution) by the wrongdoer as well as minimising future behaviour (deterrence).

Through understanding online shaming as a punishment, the goal of deterring the perceived wrongdoing emphasises on people's desire for justice (and sometimes also reflected from the goal of punishing the perceived wrongdoer). The online shaming literature suggests that people's online shaming engagement can be driven by justice-based motives and/or the aim of seeking justice (e.g., Barron et al., 2023; Chia, 2020; Hou et al., 2017; Pan, 2012). One study found that 26% of participants who engaged in online shaming were motivated by fairness concerns, as comments containing expressions related to justice (e.g., "injustice," "unfair") were commonly used (Rost et al., 2016). Indeed, to some people, online shaming may be an alternative or even a necessity to traditional justice processes (Gao, 2013; Ingraham & Reeves, 2016; Mielczarek, 2018). For example, Chang and Poon (2017) found that people who engage in online shaming (as vigilantism) tend to have low confidence in the criminal justice system and view online shaming as a more effective way to achieve justice.

Indeed, in real-life instances, people sometimes resort to shaming when formal punishment could not be sought, but online shaming has successfully led to formal punishment (Gao, 2013).

Taken together, we identified the goal of deterring the perceived wrongdoing as another motive behind the reason why people engage in online shaming. This goal is relevant to, but also distinctive from, the goal of punishing the perceived wrongdoer. Unlike the goal of punishing the perceived wrongdoer which focuses on what has been done in the past (i.e., the perceived transgression or norm violation), the goal of deterring the perceived wrongdoing is forward-thinking, aiming to deter both the perceived wrongdoer and others from committing similar transgression in the future. Therefore, we suggest that, by reinforcing social norms in a broader sense, the goal of deterring the perceived wrongdoing can reflect people's wider pursuit of social justice.

Seeking Social Acknowledgement. The third goal we identified is seeking social acknowledgement. The goal can be described as the desire for one to seek perceived social recognition and approval through engaging in online shaming. In the context of online shaming, we argue that not all sources of social feedback are evaluated in the same way. Instead, whether or not the feedback is from members who belong to the same group (i.e., an ingroup; Tajfel, 1978; Tajfel & Turner, 1979) plays a crucial role in influencing people's shaming engagement.

People experience reputational gain by participating in online shaming (e.g., Basak et al., 2019; Brady et al., 2021; Chia, 2020). Research found that people's engagement in online shaming can be predicted by their expected reputational gains and the perceived social recognition from others. For example, it was found that people's desire to "stand out from the crowd" was a key driver of their participation in online shaming ("online firestorms"; Johnen et al., 2018). As more participants engage in online shaming, people are less willing to

participate because it has become more difficult for them to stand out. However, for people who still choose to engage in shaming, their comments are found to have a more indignant tone (Johnen et al., 2018). Another research similarly found that people engaged in a higher level of aggression when they were non-anonymous than were anonymous (Rost et al., 2016). These findings were contradictory to the understanding that online shaming is mere a type of aggression or that people can become more aggressive when they are anonymous and subsumed in the crowd (Drury & Reicher, 2020; McGarty et al., 2011; Reicher, 1984). The findings imply that online shaming is not perceived as completely negative or socially unacceptable. Moreover, it is actively seen as a positive attribute when people's online shaming engagement is driven by a desire for social recognition and reputational gains.

Social recognition from one's ingroup could be especially motivating for people to engage in online shaming. According to the social identity approach (Tajfel, 1978; Tajfel & Turner, 1979), one's self concept can be derived from their belongingness to a social group (i.e., the ingroup). When one's ingroup identification is high, being part of the group could lead people to behave in a way that aligns with the ingroup norms, or the shared beliefs about what the group's values and (in)appropriate behaviours are (e.g., Terry & Hogg, 1996). Specifically on social media, online shaming can be a way to signal what one endorses as an ingroup member. This is because the norm violation that triggered online shaming is likely to have threatened their shared group norm and the underlying social identity (Brady et al., 2020). Therefore, people who engage in online shaming might view it as a moral response to restore the violated group norm. In this way, online shaming could bring social recognition and approval from one's ingroup, and fulfill people's group-identity motives.

In particular, online shaming can fulfill group-identity motives via improving the positive distinctiveness of one's ingroup. This can be achieved by supporting the ingroup and derogating the outgroup (Brady et al., 2020; Tajfel, 1981). Notably, outgroup derogation

might occur more commonly than ingroup support in the context of online shaming, due to its punitive nature (but people might also show support when their ingroup members are shamed, see Blitvich, 2022 and Tandoc et al., 2024). For example, one study found that participants who engaged in online shaming reinforced their group consensus by portraying the perceived wrongdoer as a "common enemy" (Marwick, 2021). In a series experiments conducted by Puryear (2020), when people encountered a moral dispute online, they were found to be more concerned with undermining the opponents' view rather than pursuing reputational benefits from likeminded others. Although the group nature of online shaming was acknowledged in a few research (e.g., Brady et al., 2020; Marwick, 2021; Puryear, 2020), empirical examination of these group-identity motives remains limited in the multidisciplinary literature. Future research is needed to further examine the group-based motives that could guide people's desire for social recognition and their shaming engagement, which includes the need to belong and the desire to promote and maintain a positive image of one's ingroup (Spears, 2011; Tajfel, 1982; Tajfel & Turner, 2004).

In some cases, shaming itself can be shared as a norm, which can be particularly powerful in facilitating people's engagement in online shaming. This can happen when more people started to engage in online shaming and perceived online shaming as a normative response (Brady et al., 2021; Sawaoka & Monin, 2020; Johnen et al., 2018), and/or when online shaming becomes a crucial part of one's identity (i.e., when online shaming becomes the group norm) (Murumaa-Mengel & Lott, 2023). For example, Brady et al. (2021) found that, although people's expression of outrage can be predicted by the positive feedback (e.g., number of likes) they received for the past outrage expressions, they became less sensitive to the positive feedback when there was a norm to express outrage. This suggests that the norm to shame can have a particular influence on people's engagement in online shaming. Consistently, people's intention to shame was found to be predicted by both their perceptions

of whether they should engage in shaming as a part of a community, and the extent to which other community members are performing shaming, as stronger perceived norms predicting greater intention to shame (Tandoc et al., 2024).

Taken together, we suggest that it is important to understand the goal of seeking social acknowledgement according to the desire of obtaining social recognition and approval from others, in conjunction with the group-based motives that people might have. Nonetheless, seeking social acknowledgement does not guarantee the reception of such acknowledgement. Group norms may evolve rapidly in an online environment (Postmes et al., 2000), and change the desired behaviour. Therefore, online shaming engagement is sometimes a "risky" option with potential negative consequence, such as a shaming backlash (Adkins, 2019; Jane, 2016).

Creating Change. The last goal we identified in the research is to create social change. To achieve such a goal, online shaming is used as a strategy to mobilise others towards one's cause. Online shaming driven by this goal reflects a desire to address the perceived norm violation at a group or societal level. One example is the #MeToo social movement where people publicised the allegations of sexual assaults and called out the alleged perpetrators online. When justice had not been sought, the outrage and condemnation against the alleged perpetrators not only served as a sanction, but also communicated that sexual harassment is not acceptable in the society (Leopold et al., 2021). For those who share their views on gender equality and opposition to sexual harassment, online shaming provides a way to address their group agenda.

Online shaming has also been applied to other issues with a political and/or social agenda, such as disability rights, gender equality, racial equality, as well as the power imbalance between the authority and ordinary citizens (Arancibia & Montecino, 2017; Blitvich, 2022; Gao, 2013; Kitchin et al., 2020; Leopold et al., 2021). A study examined a shaming campaign against a football club on disability discrimination found that people were

concerned with initiating changes at a group (or even societal) level (Kitchin et al., 2020). In the example, people posted not only the contents attacking the organisations and/or certain individuals, but also the contents raising others' awareness about disability rights, mobilising influential individuals for their support, and requesting club stakeholders and media to launch an official campaign on disability rights. These results suggest that online shaming can be, at least partly, driven by the goal of creating change.

When online shaming is triggered by the violation of a norm that holds particular importance to one's group identity, the shaming could be more likely to reflect the goal of creating change. In particular, we suggest that online shaming with the goal of creating change tends to be group-based. Indeed, online shaming was found to be predicted by (and sometimes defined as) shared feelings of anger and outrage following perceived injustice (Brady et al., 2020; Crockett, 2017; Johnen et al., 2018; Puryear, 2020; Sawaoka & Monin, 2018, 2020). As a group-based emotion, outrage was consistently found to play a crucial role in motivating people to effect change (collective action, e.g., Tausch et al., 2011; Thomas et al., 2012; van Zomeren et al., 2008).

Furthermore, there is evidence that participating in online shaming is associated with (or even increase) other drivers of collective action and social change, such as shared identities and perceived group efficacy (Gruber et al., 2020; Pan, 2012). Group efficacy can be defined as the belief that a group can achieve shared goals in a collective way, which was found to predict collective action (Thomas et al., 2012; van Zomeren et al., 2008). Pan (2012) found that only a small number of participants of online shaming (vigilantism) reported a feeling of being personally powerful, whereas the majority reported a sense of group efficacy as they believed the collective effort was powerful. However, Gruber et al. (2020) found that online shaming was not predicted by the perceived group efficacy, but only by the perception

of being a member within a community and the personal relevance to the issue involved in shaming.

But does online shaming lead to social change? Here the research is mixed. Online shaming can happen when someone violates a gendered norm, such as in slut-shaming (Papp et al., 2017; Papp et al., 2015). Online shaming that is triggered by this type of norm violation, followed by the portrayal of "wrongdoer" on social media and traditional media, might strengthen (rather than reduce) certain discrimination (MacPherson & Kerr, 2020; Mallén, 2016; Moore, 2016; Trottier, 2020a, 2020b). Therefore, some researchers questioned whether online shaming would always lead to social change (Brady & Crockett, 2019; Crockett, 2017; Duncan, 2020; Trottier, 2020a). It was suggested that online shaming, in general, was ethically questionable and unjustifiable. For most cases, it fails to meet one or more of the following constraints: proportionality, necessity, respect for privacy, non-abusiveness, and reintegration (Billingham & Parr, 2020).

Despite the benefits for social change, researchers also questioned whether it outweighs the ethical considerations of online shaming. Some researchers discussed the ethical and legal concerns, such as the violation of an individual's privacy and reputation rights (Aitchison & Meckled-Garcia, 2021; Bu, 2013; Chang, 2018; Cheung, 2014; Laidlaw, 2017; Ong, 2012; Oravec, 2019). Others suggested that, despite the ethical and legal concerns, online shaming might be ethically justified if it was the only way to effectively punish the wrongdoer and/or to create change (Jane, 2017; Leopold et al., 2021). For example, in a case study, Jane (2017) suggested that there was sufficient evidence to support that the online shaming practice (vigilantism) used by some feminists was effective and ethically justified, given the lack of alternative, institutional ways to intervene sexual violence.

Furthermore, we suggest that social identity could play a crucial role in shaping how people perceive online shaming. Especially, the judgement on whether online shaming is effective or appropriate for achieving change is likely shaped by social identities (see also Marwick, 2021). It was found that in the context of "slut-shaming", the identity of being a feminist played a role in shaping participants' perception on the judgement of whether shaming is justified (Papp et al., 2017). And people who self-identified as feminists were more willing to spend time with the victim and found online shaming less justified, when compared to non-feminist. Moreover, the social norms expressed can shape people's attitudes about whether online shaming is useful and justified. Chia (2020) found that people who had a greater exposure to news which favourably reported online shaming, perceived online shaming as being more socially acceptable, more useful, and less harmful, which in turn facilitated their engagement in online shaming.

By synthesising the multidisciplinary literature on online shaming, we have identified four goals that effectively integrate the literature and offer explanation to why people engage in the online shaming. The four goals are: punishing the perceived wrongdoer, deterring the perceived wrongdoing, seeking social acknowledgement, and creating change. With these goals identified, we showed that although online shaming can be understood as a punishment, it involves more than just seeking the retribution against the perceived wrongdoer. Instead, online shaming can reflect a pursuit of justice and/or a desire for social change (e.g., used as a strategy to mobilise other ingroup members to one's cause). Instead, we suggest that it is important to understand online shaming according to one's group identity.

Shifting Views on Online Shaming Over Time.

As shown in Figure 4, limited research were published on online shaming before 2016, comparing with recent years. Specifically, before 2016, there was only one research paper in sociology that described online shaming punishment with the goal of seeking social

acknowledgement (Pan, 2012), and another interdisciplinary (media and communication, and political science) paper described online shaming punishment with the goal of creating change (Gao, 2013). A closer examination of the two articles revealed that these goals (with punishment) were observed from cases of vigilantism in the Chinese context (i.e., human flesh search). Other researchers viewed online shaming in a more divided way. They either only emphasised on the punishment goal, or only the deterrence for justice, acknowledgement and/or social change without discussing it as a punishment that inflicts harm. There were two other articles in criminology and media and communication that emphasised shaming as a punishment with the goal of seeking justice (Wall & William, 2007; Wehmhoener, 2010).

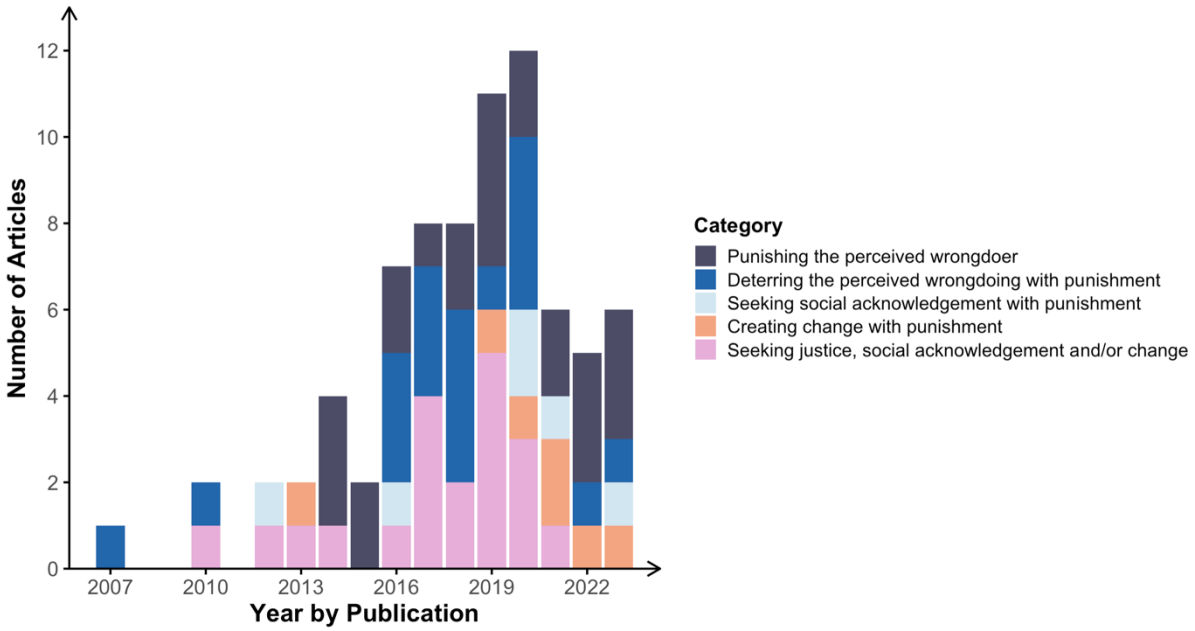
From 2016 to 2019, an increasing number of researchers started to pay attention to the phenomenon of online shaming, and the number of research papers increased substantially. However, only two articles in the fields of political science, psychology and sociology discussed or examined online shaming as a punishment with the goal of seeking acknowledgement or creating change (Oravec, 2019; Rost et al., 2016). During the same period, the views of other researchers have remained divided, though more have started to acknowledge that people could engage in shaming to seek justice, particularly in the fields of law, criminology, and psychology (e.g., Hou et al., 2017; Klonick, 2016; Loveluck, 2019; Mallén. 2016).

Since 2020, however, across disciplines, researchers have been forming a more sophisticated view on why people engage in online shaming than before. Specifically, many researchers started to see online shaming as a punishment with the goal of seeking acknowledgement and/or creating change. In comparison, the number of research papers that discussed online shaming as driving by deterrence for justice, acknowledgement and/or social change without understanding it as a punishment, started to decrease. From 2022, however, all research have started to understand online shaming as punishing behaviour that could be

driven by deterrence, acknowledgement and/or social change, which indicates a more sophisticated view on online shaming was formed. Overall, we suggest that researchers' view on online shaming shifted from being distinctively divided to having more nuances.

Figure 4

Number of Online Shaming Articles Published by Year per Category



Note. Articles were categorised based on the initial classifications that they were coded (and not coded) into. Since we were interested in how researchers understood why people engage in online shaming, the categories were based on the initial classifications about goals of online shaming, with some articles excluded for not relating to the goals. See Appendix A for details on the categorisation process.

Discussion

Online shaming has attracted widespread discussion and media attention because how ubiquitous it is and the detrimental consequences it can have (Muir et al., 2021). However, given the proliferation of the literature, there is still a lack of synthesis and consensus on why people engage in online shaming. To better answer this question, the current review provides a thematic synthesis via systematically examining the multidisciplinary literature. We brought together the diverse understandings of online shaming and integrated the multidisciplinary

literature with social psychological theories (including justice theories and the social identity approach). As a result, we propose four goals that might drive people's online shaming engagement: punishing the perceived wrongdoer, deterring perceived wrongdoing, seeking social acknowledgement, and creating change. The identification of these goals showed that, in addition to inflicting harm on the perceived wrongdoer, online shaming can also be driven by justice-focused motives and group-based motives.

The identification of the different goals has implications for how online shaming is conceptualised and theorised. We suggest that viewing online shaming as merely an individual act to harm another person should be replaced with a more sophisticated understanding. Online shaming behaviour of an individual, in isolation, may be understood as an online violence or aggression (e.g., Ge, 2020; Huffman, 2016). However, online shaming almost always happened collectively and involved one or more groups of individuals, therefore, the social basis of this behaviour needs to be acknowledged, especially when researchers examine why people engage in online shaming. The current review showed there still remains to be limited examination of the group nature of online shaming. Building on the existing research that examined the social recognition motives (Johnen et al., 2018; Rost et al., 2016), the current review further draws on the social identity approach. We propose that more research is needed to examine online shaming as a group-based behaviour.

Furthermore, the review shows that online shaming can be driven by collective goals that are shared among a group. In addition to the goal of punishing the perceived wrongdoer, online shaming also entails the goal of deterring the perceived wrongdoing (which reflects a desire for justice), seeking social acknowledgement, and/or creating change. Particularly, we suggested that the goal of seeking social acknowledgement can be understood according to the social identity approach (Tajfel, 1978; Tajfel & Turner, 1979), as people might be more concerned with the social recognition and approval of one's ingroup. Furthermore, in

identifying the goal of creating change, the current review drew a link between the online shaming literature and the collective action and social change literature that adopts a social identity approach (e.g., Thomas et al., 2012; van Zomeren et al., 2008). We suggest that future research can empirically examine these goals, especially from a social identity lens.

The finding that online shaming is often more than just a form of punishment raises the question of how it should be approached in real-life situations. Notably, one recent trend in the technology industry and academia is using machine learning techniques to detect and reduce the online comments that are considered high in toxicity (Perspective API, n.d.; Jigsaw, 2023). There were also some applications that specifically addressed online shaming (Basak et al., 2019; Surani & Mangrulkar, 2021). However, given that online shaming can entail other goals such as creating change, we suggest this approach of eliminating all "toxic" comments should be cautioned. In cases where online shaming is used by marginalised groups and/or where a formal punishment cannot be sought, reducing these types of online shaming without understanding and acknowledging what people are really demanding might cause further marginalisation and intergroup conflicts (Smith et al., 2024).

Lastly, we would like to highlight the importance of understanding online shaming within the intergroup context and in the broader societal context. Depending on the nature of the group(s) involved in online shaming, the broader, societal context could encompass political and cultural context. For example, the cancel culture that involved online shaming was found to become politicised over time in the United States. Political leaders such as Donald Trump used the narrative of "fighting cancel culture" in his electoral campaign (Fahey et al., 2022). Therefore, to understand people's shaming engagement in this case would require one to understand the intergroup context where the shaming occurs, such as the relationship between political groups in the United States. Another example is that the goal of creating change that emerged first in research about "human flesh search" (online shaming in

the Chinese context). One explanation could be that, given the barriers of engaging in offline collective actions in China (such as protests), online shaming might have played an especially important role in affecting change (Gao, 2013). Therefore, we suggest that online shaming be understood within a dynamic intergroup context that evolves alongside the broader social context.

Limitations and Future Directions

As a scoping and narrative review, we acknowledge the subjectivity in our approach to thematic analysis. We based our analysis on a selected set of key terms, which may not encapsulate every aspect of the field, given the lack of consensus on the phenomenon of online shaming and the broad range of terminology used by researchers to describe it. We also exercised judgment when refining our key terms in follow-up searches shaped by our initial research. In the analysis, while we tried to analyse what the research captured in relation to people's goals when participating in online shaming, much more could be explored when it comes to situational factors that could drive people's online shaming, as well as how situational factors shape goals and motivation. For example, researchers suggested that online expressions such as shaming can be influenced by the design of social media platforms (Brady et al., 2020) and the narrative of news media (Trottier, 2018, 2020a). These factors might also interact with people's goals to influence their online shaming engagement, which can be examined in the future research.

Additionally, although we identified the goals of online shaming based on existing research, further empirical studies need to be carried out in order to validate the identified goals. Specifically, it can be a fruitful way to use archival data to examine these goals. For example, researchers can consider observing, collecting, and analysing people's actual shaming engagement, such as the comments they made on social media. Such a method has benefits in realism and access to longitudinal data (Heng et al., 2018), which allows not only

the examination of the presence of the identified goals, but also the investigation of the emergence and progression of these goals.

Lastly, we suggest that future research should continue examining online shaming as a group-based behaviour, especially in an intergroup context. What was clear from the existing research was that the goals are shaped by intergroup dynamics in many ways, even though the behaviour is often enacting at an individual level. We suggest that future research is needed to examine whether new group identities emerge through online shaming, and if so, how the shared emerging identities might further influence people's shaming engagement. Previous research on opinion-based groups found that new social identities can develop and form via people communicating their opinions (Bliuc et al., 2007; Thomas et al., 2012), and that group-based interactions in the form of group discussion predicted people's collective actions (e.g., Smith et al., 2015; Thomas & McGarty, 2009). We suggest that perhaps online shaming as a form of group-based interaction can shape group identities and, potentially, mobilise others for social change.

Furthermore, many arguments in the online shaming literature put forward by researchers remain untested. Some researchers suggested that online shaming could be a risky option for people who choose to engage as they could face shaming backlashes and exposure of personal information as consequences (Adkins, 2019; Jane, 2016). The presence of risks raises interesting empirical questions, such as whether individuals who engage in shaming are aware of the accompanied risks but still choose to participate. Future research can also explore the perspective of those who experience or observe online shaming, building on the existing empirical studies (e.g., Sawaoka & Monin, 2018, 2020) and the understanding of online shaming as group-based behaviour. For example, it is worth examining how group memberships influence *when* and *to whom* a shaming comment is perceived as

(un)acceptable, and whether this explains why some people become observers but not engagers of online shaming.

Conclusion

To conclude, the current review made a unique contribution to the understanding of online shaming. In the review, we identified four goals from the multidisciplinary literature that could explain why people engage in online shaming. Online shaming often involves not only the goal of punishing the perceived wrongdoer, but also the goals of deterring the perceived wrongdoing, seeking social acknowledgement, and/or creating change. While literature has shown that the social basis of online shaming is gaining attention, more research is needed to examine online shaming as an intergroup behaviour. Future research should investigate these identified goals with consideration of the context where online shaming occurs. This includes the intergroup context and the broader context that encompassing the societal, political, and cultural climates.

Authorship Statement

Chapter 3 is based on a co-authored manuscript preparing for future publication:

Zhao, K., Berndsen, M., Thomas, E. & Woodyatt, L. (2025). Exploring the nature of online shaming and the progression of psychological goals: A two-phase online shaming event on Twitter. [Unpublished manuscript]. Flinders University

The candidate was the primary author of the work. Specifically, the candidate was responsible for:

- Research design,
- Data collection and analysis, and
- Manuscript writing and editing.

Co-authors provided supervision, critical feedback on study design, data collection and analysis, and results interpretation during the writing process. This work has been included in the thesis with the permission of the co-authors.

CHAPTER 3. Exploring the Nature of Online Shaming and the Progression of Psychological Goals: A Two-Phase Online Shaming Event on Twitter

Abstract

Despite the prevalence of online shaming, empirical research examining why people engage in shaming remains limited. To examine this question, we analysed a case of pandemic shaming that took place in 2020, localised in Victoria, Australia. The shaming spanned 4 days on Twitter (now X), following a doctor being called out by the Victorian health minister at a press conference. We ran topic modelling analyses on the collected tweets (N = 5,005). It was found that two distinctive communities emerged and formed, based on the different understandings on who was culpable: 1) people who shamed the doctor and expressed support for the health minister and 2) people who shamed the health minister and supported the doctor. We found that the shaming engagement of both groups were driven by the goals of punishing the perceived wrongdoer, deterring the perceived wrongdoing, seeking social acknowledgement, as well as creating change. Furthermore, as online shaming progressed, each of the two communities had unique shared norms and practices evolved, which in turn, further defined what each group represents. Especially, people who supported the doctor demonstrated system-challenging norms and engaged in collective actions, whereas people who supported the health minister showed system-justifying norms and reacted defensively to the goal pursuit by the outgroup (i.e., the doctor's supporters). These findings suggest that like online activism, online shaming may be an intergroup behaviour shaped by both collective goal pursuit and intergroup dynamics.

Introduction

On March 7, 2020, the then–health minister of Victoria, Australia, announced that the Department was working to contact individuals who might have been infected by Victoria's 11th confirmed COVID-19 case. In her announcement, the minister said she was

“flabbergasted that a doctor that has flu-like symptoms has presented to work,” adding that the GP (General Practitioner) likely contracted the virus in the United States and had treated over 70 patients since his return, including two patients in a nursing home (ABC News, 2020; Hitchick, 2020). She also suggested that the GP's conduct could be referred to the regulators, the Australian Health Practitioner Regulation Agency, for further consideration. Although the doctor was not named directly, his clinic and other details were revealed. News of the doctor quickly spread on social media, prompting some people to denounce his actions and label him as an immoral “super spreader” (see Figure 5 for the key events of this incident).

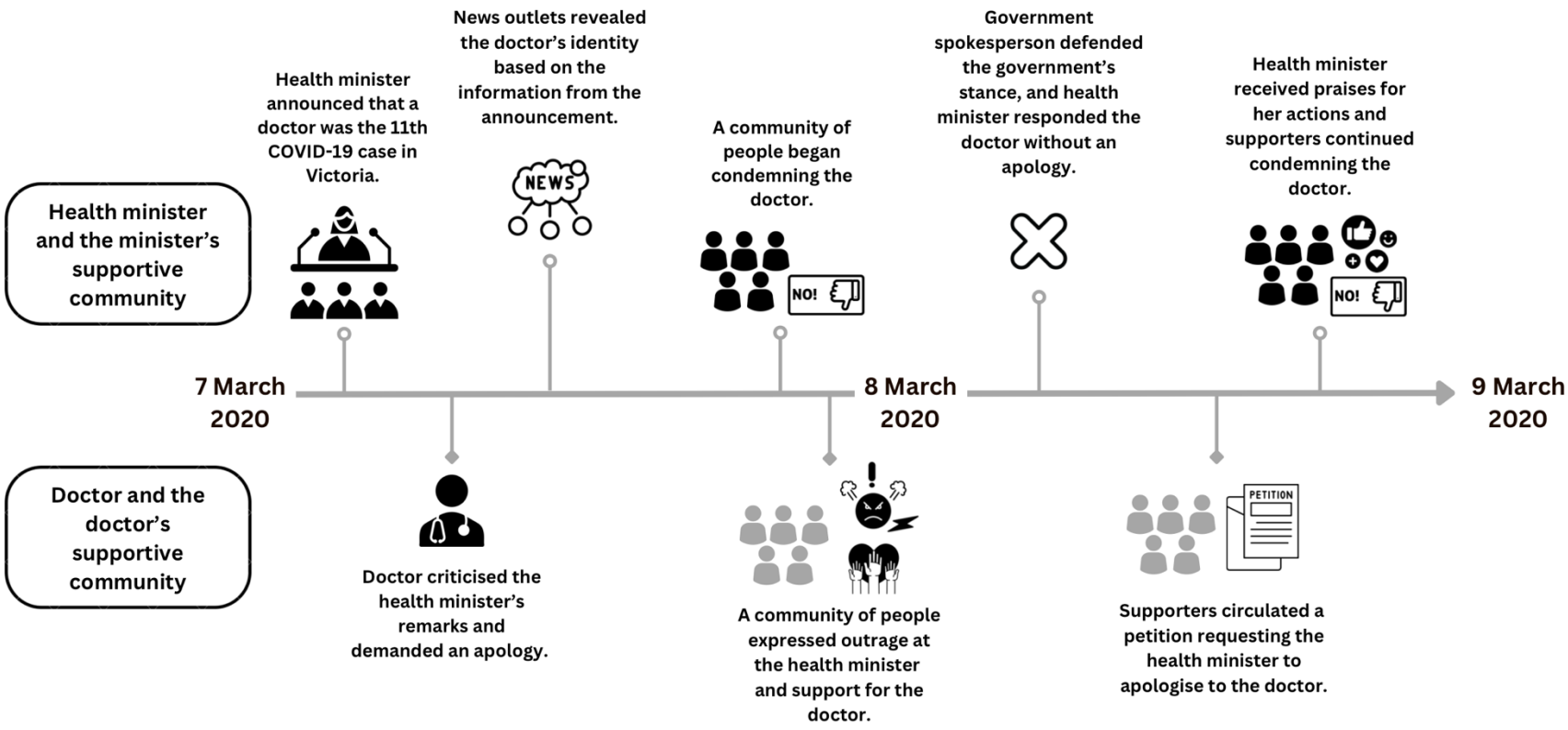
While there was no official count of how many individuals participated in shaming the doctor, the incident was significant enough to draw widespread attention. Within hours, the doctor responded on the health minister’s Facebook page, accusing her of taking “a cheap opportunity for political grandstanding” (BBC, 2020). He also demanded an apology, arguing he had not violated testing guidelines at the time. His response soon garnered attention and support from other health workers (Herbertson, 2020). Using hashtags such as *#flabbergasted* and *#flabbergaslighting*, supporters defended him and criticised health minister for unfairly condemning a doctor and undermining the confidence of the Australian medical community (Woodley, 2020). On March 8–9, 2020, the health minister responded without an apology, stating only that her thoughts were with the doctor and his patients (Swain, 2020; Chapman, 2020). This reaction further angered the doctor’s supporters, who then intensified their condemnation of the Minister while continuing to voice support for the doctor. This sentiment was soon translated into action, with a petition requesting an apology from health minister widely circulated on social media, ultimately gathering more than 11,000 signatures (Dolezal & Rose, 2020). Meanwhile, supporters of health minister not only persisted in criticising the doctor, but also extended their condemnation to other medical professionals who had sided with him.

This incident can be viewed as an example of “pandemic shaming”, a type of online shaming that are widely documented during the early stages of the COVID-19 pandemic (Cooper et al., 2023). Online shaming involves public criticism or condemnation of individuals or groups for a violation of a norm or value, usually via social media (Chapter 2). During the early stages of the pandemic, online shaming on social media surged, targeting groups such as healthcare workers, “super-spreaders”, as well as labelling individuals as “covidiot” – a word used to criticise those whose actions were seen as risking the spread of the virus (Cambridge University Press, n.d.). The widespread anxiety surrounding the then-novel virus may have heightened people's urge to protect the public good, especially in the absence of clear initial guidelines (Cooper et al., 2023).

However, despite the prevalence of online shaming, empirical research examining why people engage in this behaviour remains limited (see Chapter 2). Furthermore, as we have suggested, it is important to understand the intergroup context in which online shaming takes place. To address these gaps, this chapter analyses a specific case of online shaming directed at a doctor – and later the health minister – in a context of the early stage of the pandemic and localised in Australia. Building on the conceptualisation of online shaming and the theoretical framework outlined in the previous chapters, this chapter further explores online shaming as a group behaviour that involves both group processes and pursuit of goals, and examines whether they could explain why people engage in online shaming.

Figure 5

Timeline of the Online Shaming Event



We suggest that this specific case of online shaming can be an example of an intergroup divide between people who supported the doctor and the people who supported the health minister. As anticipated in Figure 5, it also appeared as though two distinct communities formed as the incident evolved: One group supported the health minister while shaming the doctor, whereas another group defended the doctor while shaming the health minister. Anecdotally, it also seemed to be the case that each group's stance became more defined as online shaming progressed, as shown from the different goals that the groups pursued. For example, supporters of the doctor tried to unite the medical community by circulating a petition demanding the health minister to apologise, while supporters of the health minister continued to condemn the doctor without forming a united front. To gain a deeper understanding of why people engage in pandemic shaming, the chapter examines the emergence of the distinctive groups as well as how the shared goals evolve as the events unfold.

In this research, we employ archival data to examine the group formation and shared goal(s) that might drive people's online shaming of the doctor, and subsequently, shaming of the health minister. Archival data refers to data that was originally collected for purposes other than academic research (Heng et al., 2018). Specifically, we analysed expressions on Twitter (now X) where online shaming occurred. Twitter has been one of the most popular platforms for information dissemination and discussion, especially during the early stages of the pandemic (Nanath & Joy, 2021). Unlike some other social media platforms where interactions are primarily with family, friends, or acquaintances, Twitter often facilitates engagement with strangers (Woo-Yoo & Gil-de- Zúñiga, 2014). With public profiles, users can follow and view posts without prior approval, making Twitter akin to a modern town hall or town square. It has also emerged as a

significant site for contemporary expressions of shaming, shifting these interactions from physical spaces to the digital realm (Basak et al., 2019).

Goal Pursuit in Online Shaming

Previous research has identified psychological goals that might explain why people engage in online shaming (Chapter 2). These goals were extracted from the existing literature on online shaming, and synthesised based on the potential motives that people might have when engaging in shaming. Specifically, the identified psychological goals are 1) punishing the perceived wrongdoer, 2) deterring the perceived wrongdoing, 3) seeking social acknowledgement, and 4) creating change. We explore whether these goals can be applied to the current example of pandemic shaming and used to explain why people engage in the shaming of the doctor and the health minister, respectively. Below, we first explain what each of the goals represents, followed by how they might manifest and emerge in people's online shaming expressions.

The goal of punishing the perceived wrongdoer (for short, the *punishment* goal) refers to actively seeking retribution against the norm violator (i.e., perceived wrongdoer). The goal of deterring the perceived wrongdoing (hereafter, the *deterrence* goal) refers to motives to teach the perceived wrongdoer a lesson. Although both goals address the norm violation involved in online shaming, the punishment goal addresses the norm violation that was done in the past, and the *deterrence* goal aims to change the norm violator's future behaviour (e.g., Carlsmith, 2006; Carlsmith & Darley, 2008). In a specific case of online shaming, the punishment goal can be demonstrated via inflicting the (often disproportionate) punishment on the primary norm violator, as indicated by expressions that involve harm and abusiveness (e.g., Billingham & Parr, 2020). Whereas for the deterrence goal, in teaching the norm violator and others a lesson, people

might show a tendency to reinforce the norm violation more broadly as well as express their concerns about justice and fairness (e.g., Hou et al., 2017; Rost et al., 2016), such as explaining why the norm violation was wrong and unacceptable.

Online shaming can also be driven by the goal of seeking social acknowledgement, described as one's desire to seek social recognition and approval. This goal can be understood according to research derived from the social identity approach (Tajfel, 1978; Tajfel & Turner, 1979), with people being concerned about the social recognition and approval of one's ingroup. Although online shaming typically starts with one wrongdoer being shamed, it can also involve intergroup processes. This is because the perceived norm violator who are shamed may be perceived as a specific representative or exemplar of an outgroup. Especially, shaming a representative of a broader outgroup who ostensibly shares different values can meet one's group identity-based motives, including the need to belong and to maintain a positive image of one's ingroup (Spears, 2011; Tajfel, 1982; Tajfel & Turner, 2004). Therefore, online shaming involves more than an interindividual exchange and can be a group behaviour. The goal of seeking social acknowledgement can be understood according to the group identity-based motives that reflect this idea.

Furthermore, we suggest that the group identity-based motives can be particularly relevant in understanding this specific case of pandemic shaming, in which both shaming the primary norm violator (i.e., doctor) and "shaming the shamer" (i.e., the health minister) happened concomitantly. As shown in the timeline of the online shaming event (Figure 5), there were two separate online communities that involved in this online shaming. These communities formed and pursued different (perhaps even antagonistic) interests. We suggest that the group identity-based motives could be especially salient in this case, because of the clear representation

of an ingroup-outgroup relationship. Specifically, the group identity-based motives can be expressed via both the support for the ingroup and derogation of the outgroup, with outgroup derogation being more common in online shaming (Chapter 2).

Finally, online shaming can also be driven by the related goal of creating change. People who engage in online shaming showed aims such as addressing a perceived injustice, raising awareness about a societal issue, and improving the social status of one's ingroup (e.g., Arancibia & Montecino, 2017; Blitvich, 2022; Gao, 2013; Leopold et al., 2021). Especially, the goal of creating change is likely to be presented in online shaming when it is based on a violated norm that holds a particular importance to one's group identity and perhaps relates to issues with a political/social agenda (Chapter 2). Despite that online shaming primarily reinforces the violated norm via exerting punishment (e.g., Klonick, 2016), we suggest that a goal of creating change can emerge when a group starts to focus on the advancement of their own group's interests and status (such as via showing support for their ingroup).

Through analysing a case of online shaming in-depth, the current research allows a further examination of the goals (punishing the perceived wrongdoer, deterring the perceived wrongdoing, seeking social acknowledgement, and creating change) that were previously identified from the literature (see Chapter 2). In particular, we are interested in not only the *presence* but also the *progression* of these goals. Since this case of online shaming involves people shaming different individuals (i.e., the doctor and the health minister) and the emergence of separate communities, it also provides an opportunity to examine online shaming as an intergroup behaviour. Specifically, we examine how the intergroup dynamics might change over time, along with the progression of online shaming and the evolvement of pursed goals.

Method

In the current study, we conducted correlated topic modelling (Blei & Lafferty, 2007) to analyse the corpus comprised of Twitter comments (i.e., tweets). The data collection spanned from Sunday 8 March 2020 to Thursday 12 March 2020, using Twitter's API to capture the evolution of events anticipated in Figure 5. Specifically, this online shaming incident involves both shaming of the doctor and shaming of the health minister. Notably, the doctor's refutation of the health minister's criticism appears to have been a key event that triggered the shaming of the minister. Therefore, we divided the online shaming event into two phases, based on the time when the doctor refuted the health minister's criticism. To gather as many tweets as possible, we combined two different methods of collecting tweets using the *rtweet* R package, *search_tweets* and *stream_tweets* (Kearney et al., 2022). The first method allowed us to collect tweets that have been posted in the past 6-9 days. We used *stream_tweets* to acquire a live stream of tweets. After removing irrelevant tweets, empty retweets, and tweets that were posted by news media outlets, a total of 5,005 tweets created by 2,248 unique users comprised the final corpus. Data was cleaned by removing URLs, usernames, emojis, stop words, numbers, and punctuations, to improve the accuracy of the analysis. Appendix B provides detailed information on the data-collection and data-cleaning procedures. In respect of research ethics and privacy, personal and identifiable details were removed or replaced when quoting example tweets in the current research.

Correlated Topic Modelling

Topic modelling uncovers latent themes or topics within a corpus based on the co-occurrence of words that are more likely to be associated with each other (Roberts et al., 2019). Each document (in our case, each tweet) is assumed to contain a mixture of topics, and each topic is assumed be a distribution of words that appear in the corpus (Blei & Lafferty, 2007). In

order to interpret the topics, we closely examined both the most predominant words that are highly associated with each topic (i.e., top words) and the documents that are highly associated with each topic (i.e., top tweets), followed by providing each topic a meaningful label based on the underlying theme that they represent. It is worth noting that a word or tweet can appear as a top word or top tweet for more than one topic, and that the topics can be related to each other (as assumed by the algorithm of correlated topic modelling; Blei & Lafferty, 2007).

To examine how online shaming progressed over time, we divided the tweets into two phases based on the estimated time when the doctor refuted the health minister's criticism. The first phase involved 909 tweets which were posted within the first 8 hours following the health minister's announcement, and the second phase involved 4,096 tweets that spanned for just over four days. We then ran a correlated topic modelling using the *stm* package in R for the each of the phases (Roberts et al., 2019). Before running each topic modelling analysis, we pre-processed the data using the *stm* package (Roberts et al., 2014; Roberts et al., 2016).

Model Assessment

Although topic modelling is an unsupervised machine learning technique with limited human intervention, it requires the researchers to specify the number of topics (*k*) as a priori. To determine the optimal number of topics used for the correlated topic models for the two phases, we used R package *ldatuning* (Nikita, 2016), which provides estimates for each number of *k* based on the four previously established metrics (i.e., "Arun2010", "CaoJuan2009", "Deveaud2014", "Griffiths2004"). The estimates for the first phase and the second phase were plotted in Figure 6 and Figure 7, respectively. The optimal *k* number can be determined when "CaoJuan 2009" and "Arun2010" were minimised (Arun et al., 2010; Cao et al., 2009), whereas "Griffiths2004" and "Deveaud2014" were maximised (Deveaud et al., 2014; Griffiths &

Steyvers, 2004). However, inconsistent results can be observed from the figures. As shown in Figure 6, for Phase 1 dataset, “CaoJuan2009” and “Deveaud2014” indicated that the optimal k = 3, “Griffiths2004” indicated that the optimal k = 12, whereas “Arun2010” indicated no specific optimal value k (but when k gets larger than 10-13, the model improvement was only marginal). Similarly, for Phase 2 dataset (see Figure 7), “CaoJuan2009” indicated that the optimal k = 3 (otherwise k = 15-18), “Deveaud2014” indicated optimal k = 2, “Griffiths2004” indicated that the model improvement becomes marginal when k gets larger than 12, whereas “Arun2010” indicated no specific optimal value of k.

Since the four established metrics suggested inconsistent results (particularly for Phase 1 data), we further calculated the perplexity to determine the optimal number of k for Phase 1 and Phase 2 datasets, respectively. Perplexity measures how “surprise” a model fits new data, thus a lower value suggests a model with higher generalisability. Each dataset was randomly divided into two portions, with the training set consisting of two-thirds of the tweets, and the testing set consisting of one-third of the tweets (Efron & Hastie, 2016; Hastie, et al., 2009). This random sampling procedure was repeated five times (i.e., 5-fold cross-validation) to ensure the generalisability of the results (Grimmer & Stewart, 2013; Finch et al., 2018). As shown in Figure 8, when k gets larger than 9, the model improvement becomes only marginal. Similarly, Figure 9 shows that when k gets larger than 11, the model improves marginally. Based on the results of the established metrics, perplexity, conceptual coherence (i.e., a content review of the top words associated with each topic) as well as the parsimony principle (Finch et al., 2018), we estimated a topic model of 9 topics for Phase 1, and 11 topics for Phase 2.

Figure 6

Plot of Finding the Optimal Number of Topics (k) to be Used in the Correlated Topic Model (First Phase)

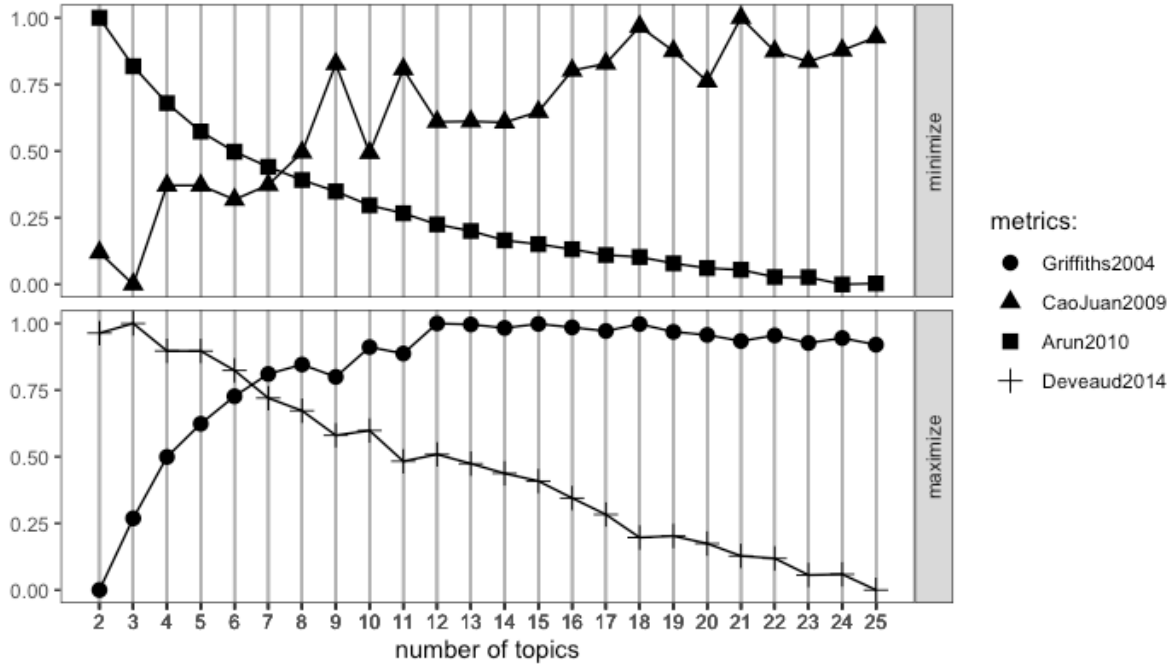


Figure 7

Plot of Finding the Optimal Number of Topics (k) to be Used in the Correlated Topic Model (Second Phase)

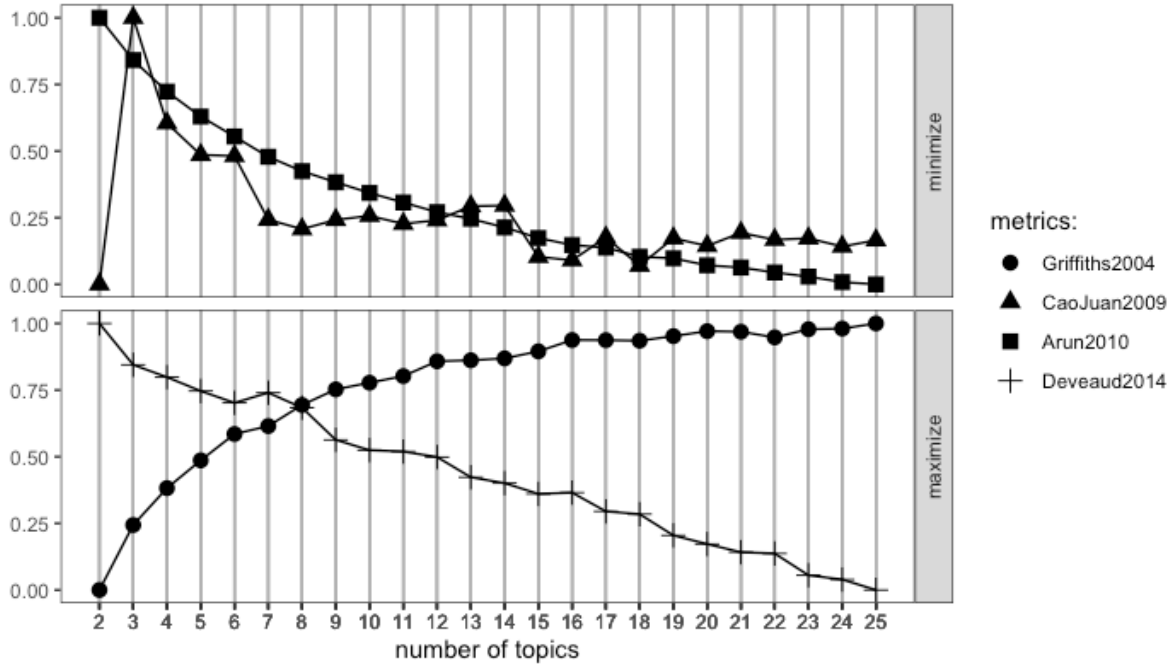


Figure 8

Five-Fold Cross-Validation of Topic Modelling (First Phase)

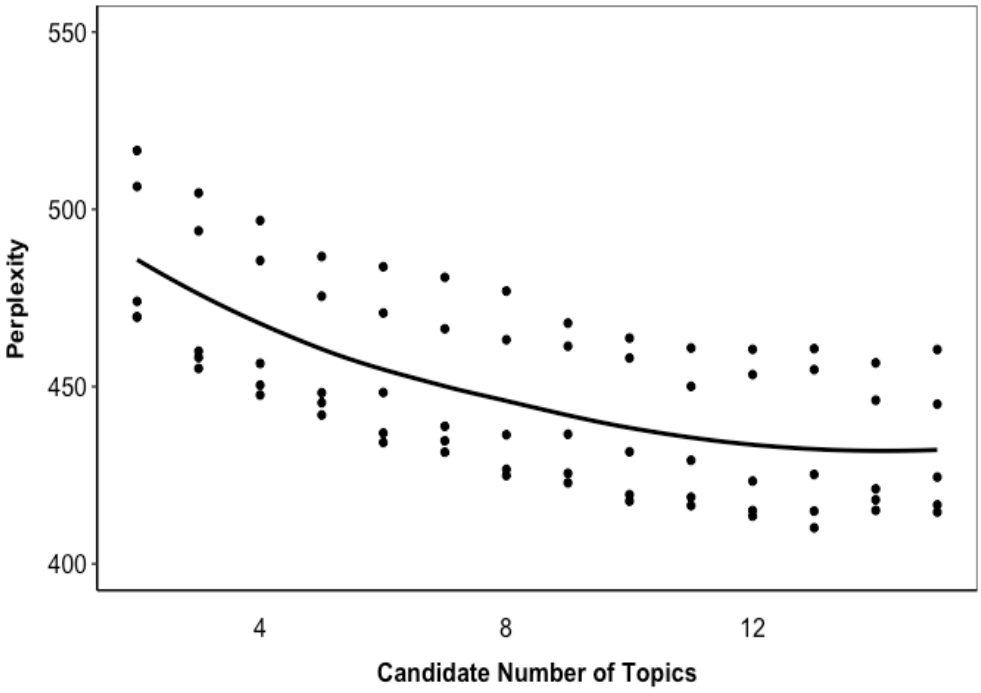
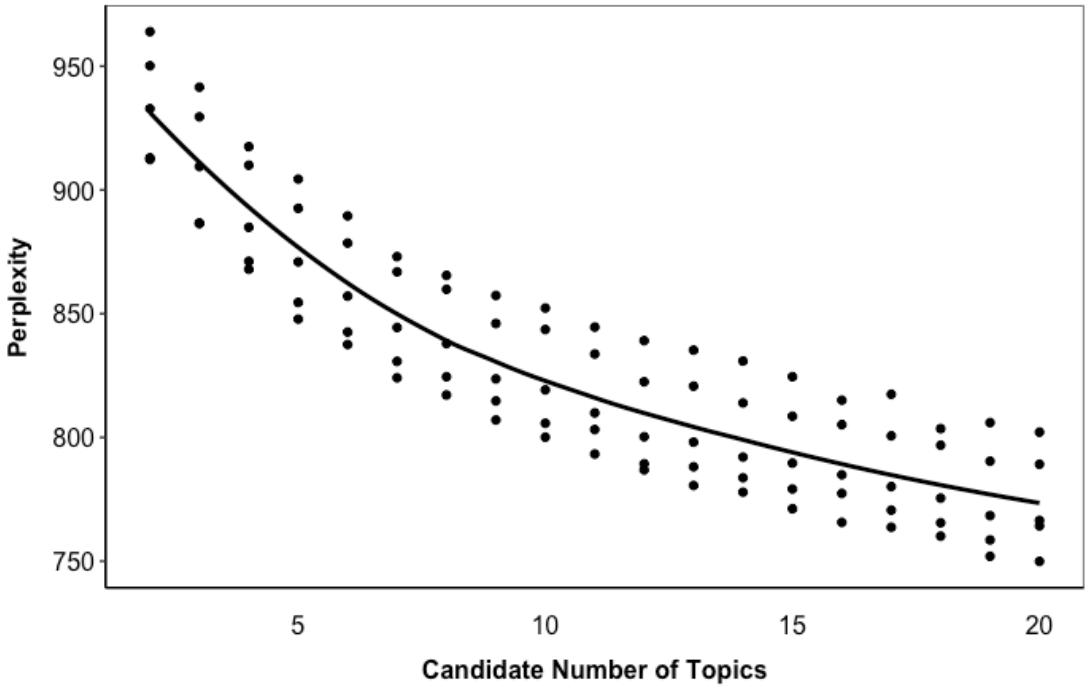


Figure 9

Five-Fold Cross-Validation of Topic Modelling (Second Phase)



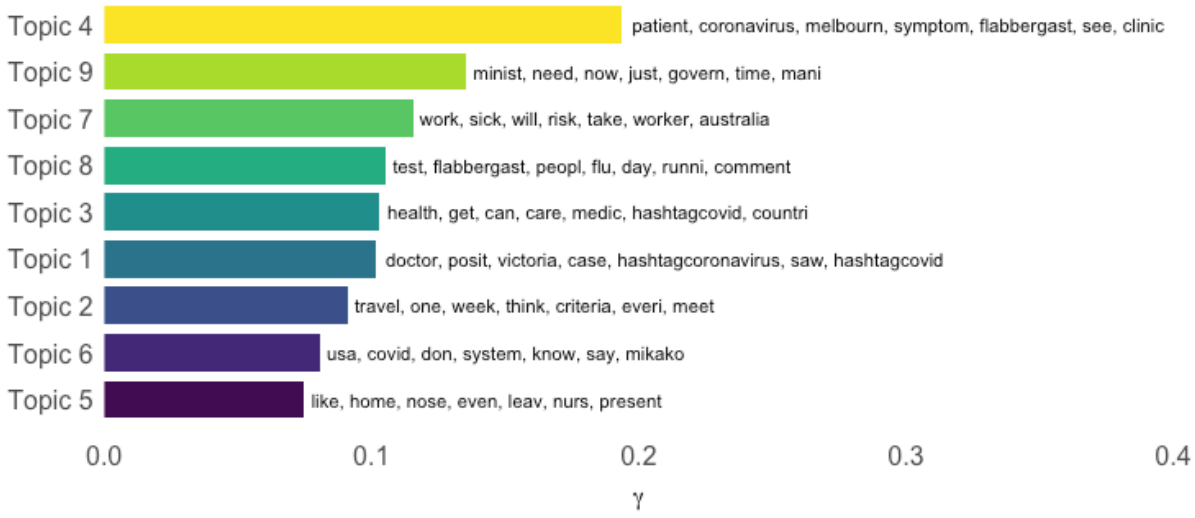
Results

Phase 1

Phase 1 involved tweets which were posted within the first 8 hours following the health minister's announcement. There were 579 unique users who commented in Phase 1 ($N = 909$), with each user contributed 1 to 18 (with an average of 1.57) tweets. The topic proportions are plotted in Figure 10, along with the words that are most associated with each topic. After closely examining the content of both the most relevant words and the most relevant tweets that are associated with each topic, we named the topics as the following: Topic 1 – Dissemination of the news (COVID-related); Topic 2 – Outrage at the health minister; Topic 3 – Disappointment at the health minister's leadership; Topic 4 – Dissemination of the news ("flabbergasted"); Topic 5 – Condemnation of the doctor; Topic 6 – Criticism of the government; Topic 7 – Pressure against health professionals; Topic 8 – Severity of the doctor's symptoms (and response); and Topic 9 – Expression of outrage at the doctor and the health minister. As shown in Figure 10, Topic 4 (Dissemination of the news["flabbergasted"]) was the most prevalent topic and had a higher proportion than the other topics, which were in general equally distributed in the data.

Figure 10

Topics by Proportion (Phase 1) with Topic Labels



Note. The x-axis represents the averaged gamma value for each topic, which represents the mean proportion with which a topic appears across all tweets (i.e., the topic prevalence). The words (in their root form) represent the top words that are most associated with each of the topics.

News Dissemination

In phase 1, a substantial proportion of the comments involve news dissemination, as captured by both Topic 1 (Dissemination of the news [COVID-related]) and Topic 4 (Dissemination of the news [“flabbergasted”]). However, these topics differed in the types of news being shared. Specifically, Topic 1 focuses more on the doctor being an infected case and a spreader of the COVID-19 virus. For example: “A doctor in Victoria is confirmed as the latest Australian to test positive for coronavirus. He had consulted approximately 70 patients over five days this past week. Melbourne GP clinic closed after doctor tests positive for #coronavirus https://...”. In contrast, Topic 4 tends to involve news reports on the health minister’s response to the doctor’s infection, emphasising what was said at the press conference. For example:

“Authorities 'absolutely flabbergasted' Melbourne doctor saw over 70 patients while infected with coronavirus https://...”. Both condemnation of the doctor and the health minister were found to accompany news dissemination, sometimes involving abusive language to criticise either the doctor or the health minister. In particular, in phase 1, condemnation of the doctor occurred more often than condemnation of the health minister. This suggests that the news dissemination was not merely about sharing information, but reflecting a goal to punish the perceived wrongdoer.

Debates About Who Was Culpable

Several topics (Topics 2, 3, 5, 8 and 9) revealed that there was a debate reflecting opposing views on who was in the wrong (culpable). It was observed that people condemned *either* the doctor or the health minister, rather than both. The debates on who was culpable can be reflected from the different expressions used for describing the doctor’s symptoms. Users who supported the doctor (and therefore condemned the health minister) described the doctor’s symptoms as only mild, using expressions such as “cold-like symptoms” and “having a runny nose”. In comparison, users who condemned the doctor described the symptoms as more severe, such as “flu-like symptoms”. The differences in expression reflect the distinct rhetorical strategies each group used, that is, the techniques they employed for effective communication (e.g., persuasion; Billig, 1996; Condor et al., 2013). Especially, in the current context of online shaming, these rhetorical strategies serve to legitimise and justify each group’s world views of who was culpable. People who condemned the doctor and those who shamed the health minister were also aware of each other’s opposing views, with sometimes people got involved into direct, intense debates with the other group, indicated by stigmatising expressions such as using swear words.

People debated on what constitutes a wrongdoing or norm violation that warrants punishment. For those who condemned the doctor, for both his behaviour and him as a person, were perceived as immoral. The doctor was perceived to have violated a moral norm: the medical profession has a duty of care, but he caused harm to his patients and to the community. For example, one user wrote "This #doctor … has been invisibly spreading #coronavirus globally. Be very careful that you are on the right side of morality". The perceived immorality of this user seemed to also form a justification for supporting the release of the doctor's personal information at a later stage: "Surely there is a public right to know re the Toorak doctor spreading the #coronavirus".

The health minister, however, was condemned for failing to be a good group leader by unfairly shaming a doctor. Some people have reframed the health minister's "flabbergasted" comment into hashtags such as *#flabbergasted* and *#flabbergaslighting* to counter its original message and posted their condemnations along with these hashtags. For example, "As the health minister I'm sure you're fully aware of how underresourced and overutilised our hospitals and general practices are. I'm #flabbergasted by your irresponsible words". In particular, it can be observed that many Twitter users who condemned the health minister were likely to be healthcare workers themselves. For example, one person expressed their outrage about the health minister's mention of the Australian Health Practitioner Regulation Agency (AHPRA) in the press conference, "…We can't seem to win. And threatening AHPRA on a doctor? - only a non doctor would fail to realise how soul and career destroying that is". These comments suggest the leader's response was perceived as not only unacceptable and disconnecting from the reality, but also akin to an outgroup member who is not part of us ("non doctor").

Taken together, we suggest that the debate on who was culpable reflected both the goal of punishing the perceived wrongdoer and deterring the perceived wrongdoing. With the punishing goal indicated by abusive and derogatory comments, and the deterring goal characterised by the attempt to reinforce norm violation that one concerned, via explaining why it was wrong and unacceptable.

Emergence of Separate Groups

The debates about who was culpable did not just function to evidence the contested nature of "wrong-doing" in shaming, it also appeared to spur the emergence of two distinct communities around the doctor and the health minister, respectively. Condemnation of the health minister seemed to be particularly driven by group-based motives, as the health minister's behaviour (i.e., shaming of the doctor) was perceived as a threat to all healthcare workers (as a group). For example, one user mentioned the health minister and wrote: "You need to publicly apologise to that GP & the entire Australian Medical community. We are on the front line risking our own lives. We are already feeling very underappreciated. Your comments were disgraceful and offensive. – Dr W, GP in Sydney." This example shows that the grievance and outrage directed at the health minister were not only shared, but also driven by a threat to identity as a medical professional specifically.

People who condemned the health minister also expressed a sense of perceived unfairness or injustice. Specifically, Topic 6 and Topic 7 included the healthcare workers' criticism of the government as well as their discussion on the pressure they faced when taking sick leave. People were outraged at the unfair condemnation on the doctor, perceiving it as an act of blaming individuals for the structural and systemic failures, especially when the healthcare system was constantly "underfunded, understaffed, and under-resourced". For example, one healthcare

worker stated that they are not seeing COVID-19 patients until the health minister apologises: "They [the COVID-19 patients] can all go to emergency department. We have to protect ourselves now, since you clearly don't have our back #flabbergaslighting". A small number of users went one step further by suggesting that the doctors should unite and take on collective action: "… GP's [*sic*] should go on strike until the Victorian health minister apologises." The indication of limited ingroup support, combined with the derogation of the outgroup, suggests that online shaming may be driven by the goal of seeking social acknowledgement. This derogation is evident in the condemnation not only of the health minister but also of the group of individuals who expressed support for the minister. However, because ingroup support was only limited in Phase 1, the goal of creating change appears to be lacking during this phase.

Online shaming of the doctor in phase 1 seems to be less driven by the perceived threat to an existing identity. Instead, people's shaming engagement seems to be driven by showing agreement with the health minister. Nevertheless, their shaming of the doctor still reflects an outgroup derogation. For example, one user indicated that "the only good thing about this totally irresponsible Toorak doctor story is that (for once) it's the 1%, rather than the poor, copping the (initial) consequences…". Unlike their counterparts, people who support the health minister lacked a salient shared identity. Nonetheless, their comments sometimes imply an antagonistic ingroup-outgroup relationship that could be motivating some users' online shaming engagement (as shown in the example tweet, identifying with the poor vs. the doctor representing the rich). Furthermore, from people who condemned the doctor, there was no mention of how to make a change at a broader community level. Instead, some of them called for further punishment against the doctor.

Taken together, in phase 1, there is evidence for the goal of punishing the perceived wrongdoer, deterring the perceived wrongdoing, and seeking social acknowledgement that can explain why people might shame either the doctor or the health minister. However, although some users who shamed the health minister mentioned collective action as a further act, the goal of creating change was not shared prevalently. On the other hand, for people who shamed the doctor, there was no indication of engagement in collective action to address broader structural issues.

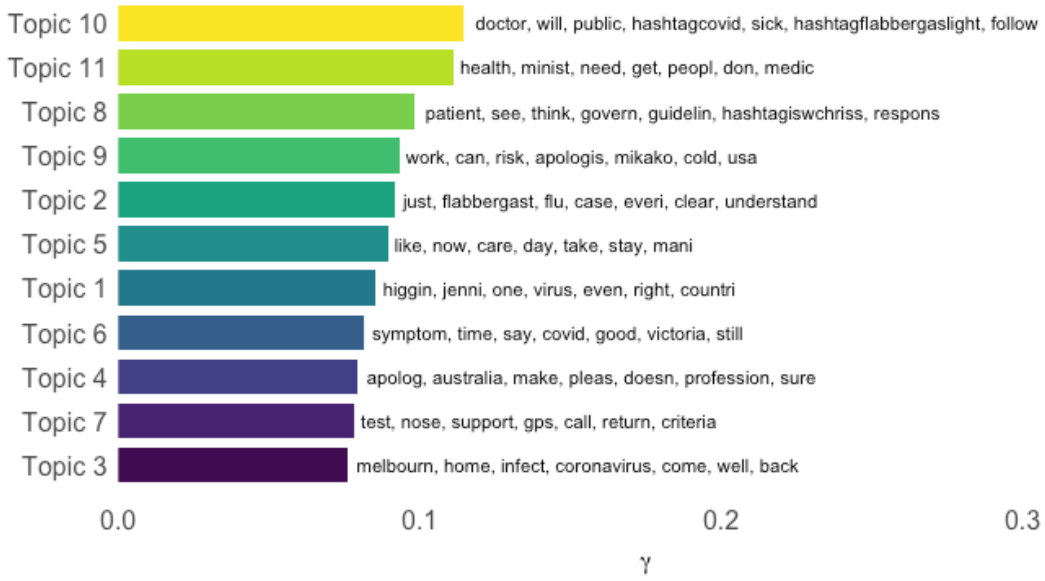
Phase 2

Phase 2 started with the doctor's response of refuting the health minister's call-out and requesting an apology. Phase 2 involved tweets that spanned for just over four days. There were 1841 unique users who commented in Phase 2 ($N = 4096$), with each user contributed 1 to 55 (with an average of 2.22) tweets. The topic proportions are plotted in Figure 11, along with the words that are most associated with each topic. After closely examining the content of both the most relevant words and the most relevant tweets that are associated with each topic, we named the topics as the following: Topic 1 – Moralisation of the doctor and the health minister's behaviour; Topic 2 – Outrage at the health minister's comment; Topic 3 – Information sharing and sensemaking; Topic 4 – Demanding apology from the health minister; Topic 5 – Debate on the correct response of healthcare workers; Topic 6 – Debate on the correct response of the doctor; Topic 7 – Calling support for the medical community; Topic 8 – (Collective) goals being further defined; Topic 9 – Identifying the work culture; Topic 10 – Intensified condemnation of the health minister or the doctor(s); Topic 11 – Condemnation of the health minister (for a failed leadership). Different from Phase 1, the information sharing (Topic 3) was no longer the most

prevalent topic in Phase 2. Rather, the topics were in general equally distributed in the data (as shown in Figure 11).

Figure 11

Topics by Proportion (Phase 2) with Top Words Contribute to Each Topic



Note. The x-axis represents the averaged gamma value for each topic, which represents the mean proportion with which a topic appears across all tweets (i.e., the topic prevalence). The words (in their root form) represent the top words that are most associated with each of the topics.

Intensified Condemnation and Outgroup Derogation

The condemnations of the doctor and the health minister further intensified in Phase 2, as reflected in the Topic 2 (Outrage at the health minister's comments), 10 (Intensified condemnation of the health minister or the doctor[s]), and 11 (Condemnation of the health minister [for a failed leadership]). Consistent with Phase 1, people condemned either the doctor or the health minister, and debated who was to blame. Despite that the doctor was refuting the health minister's criticism and explaining that he had followed the testing guidelines, only a few

users changed their mind on the issue and indicated that they no longer agreed with the health minister's criticism. The majority who continued shaming remained committed to their prior positions, posting comments that were even more abusive and stigmatising than in the previous phase.

In particular, it can be observed that people who shamed the health minister went beyond condemning her wrongdoing and extended to attacking her as a person, suggesting the shaming has become more stigmatising in phase 2 (Braithwaite, 1989). In these attacks, people expressed their disappointment with the health minister and further emphasised that she belongs to an outgroup (i.e., politicians). For example, one user tweeted: "(…) Shame on you. You politicians stand in ignorance of science and ask GPs to work so hard. Then you grandstand by releasing private medical details. You should resign in shame, to convey your apologies (…)". And in another example: "(…) Never admit you're wrong, never apologise - @Username[health minister's account] illustrates why doctors are an admired, respected & trusted profession - and why politicians are NOT! Shame on her & the Govt".

The above examples showed that an antagonistic ingroup-outgroup relationship further developed in the current phase, suggesting that online shaming has continued to be driven by derogation of the outgroup. Especially in the current phase, not only was the health minister derogated, but other outgroup members who expressed support for the health minister were also derogated. For example, one user responded negatively to someone who supported the health minister: "… How very ignorant of you [name of the minister's supporter]! Shame. She's doing everything she can? Really? Wow. You sound just as silly as her". It can be demonstrated that, for the doctor's supporters, the perceived outgroup has become more clearly defined along the evolution of online shaming. Previously, while the health minister was excluded from the

ingroup (by being referred to as a "non-doctor"), the perceived outgroup remained loosely defined. However, as online shaming progressed, a clearly defined outgroup emerged, consisting of the health minister and her supporters.

Outgroup derogation also persisted among the supporters of the health minister, often involving stigmatising comments. Especially, the doctor's request for an apology was perceived as undeserving: "(…) dont [*sic*] apologize to [the doctor's daughter's name] idiot doctor father, he should be prosecuted for endangering public health, how is he even allowed to practice medicine at all". Tweets condemning the doctor in the current phase can be abusive and involves personal attacks, including references made to his identity of being a father of a celebrity, his gender, as well as his socioeconomic status. People also condemned the healthcare workers who showed support for the doctor, for instance: "(…) I have a lot of respect for many individual GPs, but often when the profession talks together, it feels like they live on another planet". Therefore, online shaming (and especially, the outgroup derogation involved in shaming) was used by health minister's supporters as a tactic for the ingroup to counter the outgroup's group-based expressions.

Taken together, online shaming in phase 2 continued to be driven by the goal of punishing the perceived wrongdoer and seeking social acknowledgement. Furthermore, the salience of outgroup derogation involved in online shaming suggests that it can be understood as a group behaviour that is shaped by intergroup interactions. We further unpack this idea by explaining how the outgroup's response can shape people's perceptions of unfairness as well as driving people's shaming engagement.

Continuance and Emergence of Perceived Unfairness

In phase 1, we found that supporters of the doctor perceived the shaming of the doctor as unfair, perceiving it as an act of blaming individuals for systemic issues. The perceived unfairness became even more salient in phase 2 (as shown in Topics 1, 5, and 6). Specifically, the doctor's supporters continued to question the ongoing shaming of the doctor, with some emphasised the role of media in contributing to the shaming of the doctor. It was suggested that there was no justification in naming the doctor's celebrity daughter, and there were "so much sensationalism" in the reporting of this incident. For example, one user said: "I am annoyed at the release of the name, photograph, and personal details of the Toorak doctor being reprimanded by the minister and health officials in their announcements. This isn't fair and in this environment it's reckless if not deliberately provocative - so who did it?" This perceived active engagement and contribution of the news media seemed to have further intensified the sense of unfairness shared by the doctor's supportive community. As a result, unlike the previous phase, where the sense of unfairness had only begun to emerge, in the current phase, supporters of the doctor transitioned to having an explicit, collective goal (i.e., to request an apology).

This collective demand for the health minister to apologise appeared to have also cultivated a sense of unfairness felt by her supporters, which was lacking at the beginning of the online shaming event. The health minister's supporters seemed to perceive the collective request for the health minister to apologise as going too far, given that the supporters believed the minister did the right thing to call out the doctor. For example, one user posted: "(…) Ridiculous! Dr [Surname of the doctor] needs to be apologising to the 70 people he had contact with when treating them. But of course, people side with the rich male doctor rather than the female health minister. (…)". And "only a white privileged male would have the gall to demand an apology

and not even have the decency to extend an apology to the patients and families he has put at risk." Furthermore, people used hashtags such as *#Dr[name of the doctor]SaySorry* and *#IStandWith[name of the health minister]* when engaging in shaming, as a response against the doctor's and his supporters' request for an apology. These tweets and hashtags showed that people's shaming engagement can be driven by their perceived unfairness, which can be elicited and shaped by the collective goals pursued by the outgroup.

In sum, the sense of unfairness expressed by each group as well as the debates between the groups reflect that online shaming has continued to be driven by the goal of punishing the perceived wrongdoer as well as the goal of deterring the perceived wrongdoing. Furthermore, as an ongoing process, shaming was found to be shaped and sustained by the active participation of the news media as well as the outgroup's identity-based expressions. This shows that online shaming can be driven by group-based expressions and goals pursued by the outgroup, demonstrating that online shaming can be a group process that evolves constantly.

Expression of Collective Goals

Consistent with the prevalently shared sense of unfairness, the doctor's supporters asked the health minister to apologise for her wrongdoing, or even "resign in shame" to convey apologies. Tweets that shared petitions emerged in the current phase, showing a goal of creating change: "Anybody who has followed the story would be "flabbergasted" about the remarks (…) He deserves a public apology from the Victorian health minister. Please sign this petition requesting that she do the right thing. https:// (…)". And another example, "… An apology from [Name of the health minister] MP for undermining confidence in the medical profession - Sign the Petition! https://(...) via @ChangeAUS". Particularly, the doctor's supporters (who appeared to be healthcare workers) also advocated for more support from the health minister and the

government in responding to the pandemic. This includes more transparency from the government, setting up a channel for virtual consultations between doctors and patients, as well as providing healthcare workers with enough protective gears.

The doctor's supporters clearly proposed what they wanted to achieve as a group. As shown from the associated topics (Topic 7, 8 and 9), the doctor's supporters wanted a public apology from the minister, not only for public shaming the doctor, but also for undermining people's confidence in the medical profession. Consistently, the health minister was condemned for not only blaming a single doctor, but also for ignoring "the bigger issue" with the medical system, such as the lack of clear clinical guidelines for COVID-19, lack of funding, and the lack of redundancy in staff numbers. Therefore, although the online shaming of the health minister initially aimed at seeking justice for the doctor, over time the medical community united not only for supporting the doctor but also for supporting the wider medical community, showing a pursuit of addressing systemic issues and creating change.

People who supported the health minister, however, showed a different collective goal to those who supported the doctor. Unlike those who supported the doctor, although the health minister's supporters started to share a sense of unfairness in the current phase, their sense of unfairness seemed to be more reactive and based on the shaming responses of the healthcare workers. People suggested that the healthcare workers should stop supporting each other or rallying around the doctor who has clearly made a mistake. As one user put forward: "… I agree with [the health minister] Someone standing up for patients and not pandering to the media and PC [political correctness] rubbish. I'd hate to go to the GP for a sore back and end up with the flu,let [*sic*] alone coronavirus. It's irresponsible not to name the GP for all who was in contact".

Tweets like this showed that consistent with phase 1, people defended the health minister by expressing agreement with her action to publicly condemn the doctor.

The different collective goals pursued by the two groups align with research that distinguishes between system-challenging and system-supporting collective actions (e.g., Jost et al., 2017; Osborne et al., 2019). Especially, the doctor's supporters appeared to challenge the existing medical system by emphasising the systemic issues faced by the healthcare workers. In contrast, the health minister's supporters defended the health minister's act of shaming the doctor, showing a goal that aims to support, rather than challenge, the existing system. Additionally, only the group who supported the doctor, but not the group who supported the health minister, engaged in collective action by signing and sharing petitions.

Taken together, in the current phase, people's expressions from both communities (either supporting the doctor or the health minister) continued be driven by the goal of punishing the perceived wrongdoer, deterring the perceived wrongdoing, as well as seeking social acknowledgement. The goal of punishing the perceived wrongdoer and the goal of seeking social acknowledgement became more salient, as supported by the intensified condemnation and outgroup derogation. However, when it comes to the goal of creating change, although both groups demonstrated collective goals, their goals differed, as the doctor's supporters aimed at advancing their group agenda, and the health minister's supporters aimed at defending the minister's shaming act, with only the doctor's supporters transitioned in engaging in collective actions.

Discussion

The current study is amongst the first investigations of the psychological goals that drive online shaming in everyday social interaction. To examine the relevance of the goals of online

shaming that were identified in the scoping review (Chapter 2), Chapter 3 analysed an online shaming event that occurred at the beginning of the COVID-19 pandemic. Overall, the analysis of the event showed that all the proposed goals were supported, including the goal of punishing the perceived wrongdoer, deterring the perceived wrongdoer, seeking social acknowledgement, as well as creating change. Together, our findings support the idea that online shaming can be guided by diverse motives, which can be shaped and influenced by intergroup processes.

In keeping with our proposition that online shaming frequently involves intergroup interaction, the results showed that two groups emerged and formed as the shaming event unfolded. These groups differed in their views on who was in the wrong, with one group condemning a doctor who was perceived to have carelessly infected patients, and the other a health minister who was perceived to have unfairly called out the doctor. Over time, both intragroup polarisation and intergroup dynamics were found to be at play in online shaming, with people interacted with other likeminded individuals (their ingroup) as well as those who opposed them (the outgroup). Specifically, we found that outgroup derogation was prevalent among members of both groups since the beginning of the online shaming, and further intensified as the event progressed. The collective goals that drove each group to engage in shaming also became more salient over time.

As online shaming progressed, the two groups also began to differ in the norms and practices that evolved. Specifically, the group who shamed the health minister (or the doctor's supporters) tended to be comprised of medical professionals. Through shaming the health minister, the medical professionals united to express support for the doctor as well as for other healthcare workers (i.e., their ingroup). Especially, our analysis showed that although there was a pre-existing, shared identity presented for people who condemned the health minister, their

group identity became more defined based on shared ideas. As in the eyes of people who shamed the health minister, the minister's call-out of the doctor was wrong – an act of blaming individuals for the structural and systemic failures in the healthcare sector. With the group agenda became articulated (i.e., to challenge the existing status quo), the doctor's supporters demonstrated a shared, collective goal to create change, which ultimately transitioned into actual collective actions (i.e., signing and sharing petitions that ask the health minister to apologise and/or resign).

Consistent with research that proposed norms and identities can form via communications (Smith et al., 2015), we found that online shaming can be understood as a group discussion that allows shared norms and identities to form. Especially, Smith et al. suggested that when people experience a normative conflict between "how things are" and "how things should be like", they can communicate and negotiate injunctive norms (i.e., how one should behave in the given context; Cialdini et al., 1991) via group discussion. This process can then form a basis for new group identities, which can initiate or drive a social movement forward (Smith et al., 2015). Indeed, in our analysis, we found that people's shared perceptions on who was in the wrong, as well as the related justifications and expressions of emotions, became a basis for the shared norms to form, which in turn, further defined what the group represents.

Similarly, shared norms and practices also evolved among the health minister's supporters, though based on a different understanding on who was in the wrong. Instead of viewing the health minister's act of calling out the doctor as unfair, minister's supporters defended the health minister's act and reacted defensively to the goal-pursuit of the doctor's supporters. However, instead of showing ingroup support, the health minister's supporters extended the shaming to the doctor's supporters. They also demonstrated a goal that was more

defensive and tended to maintain or justify the existing status quo (i.e., system-justifying; Jost et al., 2017). Consistently, although the health minister's supporters also demonstrated a collective goal, they *did not* engage in collective actions.

Furthermore, we would like to emphasise the differences between the shared identities that underlie the groups, which could be especially relevant in the context of pandemic shaming. Since the doctor's supporters seemed to be mainly comprised of fellow doctors and other healthcare workers, there was a well-defined, existing ingroup for the doctor that was based on the doctor's profession. These healthcare workers may have also experienced a long-standing, shared grievance about the chronically under-resourced healthcare system (Willis et al. 2021), a sentiment that might have further intensified due to the uncertainty imposed by COVID-19 (Rafferty et al., 2021). However, for the health minister, the ingroup was formed over time and less well-defined, comprised a higher proportion of members of the public more generally. Furthermore, the group who supported the health minister formed based on their agreement with the health minister, lacking such a long-standing, shared sense of grievance. We suggest that due to the healthcare workers' awareness of the structural issues within the sector, the doctor's supporters might have an identity this is more politicised than the health minister's supporters. This difference could make the doctor's supporters more ready to engage in collective action than those who support the health minister (e.g., van Zomeren et al., 2008). This further explains why it was the doctor's supporters, rather than the health minister's supporters, engaged in collective actions.

It is worth noting, however, despite the lack of a long-standing grievance, people who supported the health minister started to form a sense of unfairness as online shaming progressed. We suggest this cultivated sense of unfairness among the health minister's supporters provides

crucial evidence that online shaming is a *dynamic intergroup process*. Unlike the healthcare workers who started to express a sense of unfairness since the start of the event, the health minister's supporters' sense of unfairness was cultivated by the outgroup's shaming expressions (involving both the condemnation of the health minister and the support for the doctor). Furthermore, as a response to the outgroup's goal to challenge the existing the healthcare system, the health minister's supporters formed a goal to defend the response of the authorities, suggesting they wanted to maintain the existing status quo between the authorities and the healthcare workers. Hence, shaming as a dynamic intergroup process is also reflected in how the collective goals demonstrated by the health minister's supporters were shaped by those pursued by the outgroup.

Taken together, we have provided one of the first empirical evidence that online shaming is an intergroup behaviour, which can be driven by both intergroup dynamics and collective goals that the groups pursue. We found that through engaging in online shaming, each of the two groups (either people who shamed the doctor or the health minister) had unique shared norms and practices evolved, which in turn, further defined what each group represents. Especially, people who shamed the health minister demonstrated a system-challenging norm and engaged in actual collective actions, whereas people who shamed the doctor demonstrated a system-justifying norm to maintain the status quo. This suggests that online shaming is analogous to online activism and can be used for groups both creating and opposing changes.

These findings have expanded our understanding on online shaming and emphasised the intergroup nature of this phenomenon. By providing evidence for the multiple goals that can drive one to shame (which include not only to punish someone but also to create change), we showed how online shaming can be analogous to online activism. The collective goals that drove

people to shame were shaped by intragroup interactions as well as intergroup dynamics, suggesting that online shaming can involve groups with different (or even opposing) agendas. Especially, consistent with research on that apply system-justification perspective to collective action (Jost et al., 2017; Osborne et al., 2019), we found that online shaming can either be used to drive change and challenge the status quo, or to support the status quo and defend against change. This extends the discussion on the dialectical nature of collective action (Thomas & Osborne, 2022; Osborne et al., 2019) by showing how everyday interactions on social media can embody dialectical forms of online activism.

It is worth noting that this study is limited to a single case of pandemic shaming. This restricts the generalisability and direct applicability of its findings to other practical contexts. Nonetheless, this study not only offers valuable insights into the context of other public health and environmental crises, but also the broader public debates about the online shaming phenomenon. For example, there are ongoing debates on whether social change can be achieved via online shaming (e.g., Vogels, 2022; Vogels et al., 2021). We suggest that similar to collective actions in general, online shaming might not always entail a progressive or liberal goal that benefits disadvantaged groups but can also involve a reactionary or conservative goal that seeks to maintain the system (e.g., Becker, 2020; Thomas & Osborne, 2022). Furthermore, because of the intergroup nature of online shaming, the cause of an online shaming might be viewed as positive and benefiting the society merely by some people (e.g., who support the group's agenda), but not by the others (e.g., who oppose the group's agenda).

Limitations and Future Directions

The present study has several methodological limitations. Firstly, although topic modelling allowed an in-depth study on the Twitter comments people posted and replied, it

lacked an analysis on the other responses that people might have received (e.g., “likes” and “shares”). These other responses could offer a further opportunity to explore of some of the goals underlying online shaming. Specifically, responses such as “likes” and “retweets” (“shares” on Twitter) could indicate a social acknowledgement that people received from the others, as in certain cases were found to increase people’s likelihood of future outrage expressions (Brady et al., 2021).

Another limitation is that the source of data was restricted to the comments that people posted on Twitter. However, shaming of the doctor and shaming of the health minister also took place on other social media platforms, such as on Facebook where the doctor refuted the health minister’s criticism (Swain, 2020). Previous research suggests that people may behave differently across platforms due to the varying affordances offered by each (Oz et al., 2018; Waterloo et al., 2018). Therefore, it is plausible that the development and expression of online shaming could vary across platforms. For example, the design of Facebook’s interactive emojis (i.e., Reactions) allows users to engage in a more nuanced yet low-effort way, potentially complementing or serving as an alternative to comments (C. Kim & Yang, 2017; Masullo, 2022). Consistent with these limitations, we suggest that future research could examine the phenomenon of online shaming via analysing multiple sources of data. This could involve cross-platform analysis and the combination of comments with other engagement metrics, such as likes and shares.

Although not being a focus of the current analysis, there is also evidence that online shaming can be shaped by certain social actors, such as the press and news media. In the current shaming event, the condemnation of the doctor initially appeared at the press conference led by the health minister’s (as indicated by the “flabbergasted” comment), followed by news media

that publicised the doctor's personal information. During shaming, the news media further publicised more detail of the doctor (e.g., the doctor's daughter being a celebrity), which seemed to shape (and perhaps also partly guide) people's ongoing condemnation of the doctor. For those who supported the doctor, the release of doctor's personal information might have intensified their perceived unfairness. These results are consistent with research emphasising the role of press and news media in influencing public perceptions of online shaming (Muir et al., 2021; Trottier, 2020b). We suggest that future research further explore the interplays between the press or news media and intergroup dynamics in shaping people's online shaming engagement.

Future research can also explore how the role of leader can influence people's online shaming behaviour. In the current case of online shaming, the health minister as a leader might have influenced both groups who engaged in online shaming. It was observed that for people who shamed the health minister, as the event progressed, people expressed more disappointment with the health minister's leadership and further emphasised that the minister was akin to an outgroup member (i.e., "a politician"). For people who shamed the doctor, however, they formed a group based on their agreement with the leader and shamed the doctor following the leader's call-out of the doctor. This suggests that the role of leader can influence or guide people's online shaming behaviour in different ways. Specifically, future research can explicitly examine how having a shared identity with the leader can shape people's online shaming engagement.

Conclusion

In this study, we have provided one of the first empirical evidence that online shaming is an intergroup behaviour, which can be driven by both intergroup dynamics and collective goals that the groups pursue. Especially, through analysing a case of pandemic shaming that involved the formation of two different groups, we found support for the goals of punishing the perceived

wrongdoer, deterring the perceived wrongdoing, seeking social acknowledgement, as well as creating change. We also found that through engaging in online shaming, each of the two groups had unique shared norms and practices evolved, which in turn, further defined what each group represents. Specifically, one group demonstrated a system-challenging norm and engaged in actual collective actions, whereas the other group showed a system-justifying norm to maintain the status quo. We suggest these findings showed that online shaming can be used by groups for either supporting or opposing social change, in ways that are analogous to online activism.

Authorship Statement

Chapter 4 is based on a co-authored manuscript preparing for future publication:

Zhao, K., Berndsen, M., & Woodyatt, L. (2025). Online shaming as group-based punishment: (Dis)engagement through leader's mobilisation [Unpublished manuscript]. Flinders University

The candidate was the primary author of the work. Specifically, the candidate was responsible for:

- Research design,
- Data collection and analysis, and
- Manuscript writing and editing.

Co-authors provided supervision, critical feedback on study design, data collection and analysis, and results interpretation during the writing process. This work has been included in the thesis with the permission of the co-authors.

CHAPTER 4. Online Shaming as Group-Based Punishment: (Dis)Engagement Through Leader's Mobilisation

Abstract

Online shaming has been widely referred to as a punishment. However, there is still little empirical research that explicitly examines this idea. Based on the understanding that online shaming can be understood as group behaviour driven by the pursuit of collective goals, we examined online shaming behaviour as a group-based punishment in two experimental studies. Specifically, through a lens of leadership-followership dynamic, we used two paradigms where the collective goal of online shaming was to elicit certain types of prosocial behaviour. In Study 3, participants (N = 174) were presented to either a mobilising identity leader (who belongs to an ingroup and presents a punitive norm with a noble goal) or a non-mobilising leader (who belongs to an outgroup and presents no information about the norm or goal), to examine whether participants would be mobilised by the identity leader to shame others. In Study 4, to further investigate the leader's mobilisation, we disentangled the manipulations used in Study 3 using a three-way factorial design (N = 406; manipulating the leader's group membership, the presence of leader's norm and goal, and whether the shaming punishment is aggregated). Across two studies, we found that people generally did not prefer using online shaming for punishment or deterrence. Furthermore, in Study 4, people's perceived appropriateness of shaming comments and their behaviour intention to shame were shaped by the identity leader, only when the identity was salient and when the leader's punitive norm and noble goal were not explicitly imposed. Therefore, we suggest that online shaming is often accompanied with consequences that may counteract the collective goal, even if it is a noble one.

Introduction

In December 2013, before boarding an 11-hour flight to South Africa, Justine Sacco posted a joke on Twitter: "Going to Africa. Hope I don't get AIDS. Just kidding. I'm white!" (Ronson, 2016). With only 170 followers on her Twitter account, Sacco would not have predicted that a massive, worldwide online shaming was brewing during her flight (Ronson, 2016; TED, 2015). A massive amount of criticism soon flooded into her account, calling her a racist. In a matter of hours, the hashtag *#HasJustineLandedYet* went viral and became the top trending topic on Twitter. By the time she landed she had lost her job. Due to the massive online shaming, Sacco experienced severe reputational damage, enduring emotional distress, and harassment, including death threats (Ronson, 2016). But why do people engage in online shaming actions that create so much harm to one person? This chapter explores that question in depth.

One explanation to this question is that people could be mobilised to punish others by a leader who makes social identities salient to their followers, where people ignore the harm and engaging in online shaming behaviour, in pursuit of groups norms and values. A journalist, Sam Biddle, seemed to have played a role of leader in the case of Sacco. He later wrote in another blog (Biddle, 2014): "As soon as I saw the tweet, I posted it. I barely needed to write anything to go with it: This woman's job was carefully managing the words of a large tech-media conglomerate, and she'd worded something terribly." His sharing put Sacco under the public spotlight, leading thousands of people to follow his lead in shaming Justine (Ronson, 2016). Especially, the shaming that he led can be understood as a form of third-party punishment (Fehr & Fischbacher, 2004), where people who shamed Justine were not directly harmed by Justine's violation nor seemed to receive any direct benefit from the shaming. Drawing on the insights from the case of Justine Sacco, we investigate online shaming behaviour as a *group-based*

punishment and examine it through *a lens of leadership-followership dynamic*. Particularly, using a novel experimental paradigm, we explore how social-identities may be involved in emerging leadership-followership dynamics that mobilise support and use of shaming behaviours.

Online Shaming as Group-Based Punishment Guided by Leader

Online shaming can be understood according to the social identity approach and be conceptualised as a group behaviour (see Chapter 1 for discussion). This approach proposes that an individual may define oneself in terms of a shared social identity, which can be meaningful and important to the individual's self-concept as well as how they understand the social world (Haslam et al., 2010; Tajfel & Turner, 2004; Turner, 1982/2010). Individuals have multiple group identities that can become salient depending on the relevant context and situation (Spears, 2021), such as being a member of a professional group at a work-related conference or of a national group at watching the Olympics. When a social identity becomes salient, an individual who identifies with the group can perceive themself as interchangeable with other group members (i.e., the process of depersonalisation; Turner, 1982/2010). Accordingly, the individual can internalise the group values and norms, and look up to other ingroup members in terms of how to behave in the situation (Spears, 2021; Turner et al., 1987).

In the case of Justine Sacco, the context was particularly relevant for a social identity to become salient, due to the leader (i.e., Sam Biddle)'s reinterpretation of Sacco's tweet. Sacco later recalled that her tweet was originally intended to highlight the obliviousness of some Western perspectives toward issues in Africa, serving as a mockery of white privilege (Ronson, 2016). However, Sam Biddle interpreted it in the opposite way. In his retweet and the comments that followed, Sacco's tweet was perceived as blatantly offensive, leading to her being portrayed as a racist. In this context Biddle's comments mobilise online shaming as a form of group-based

punishment that is exerted by a group of people who were against racism and wealthy, white privilege. The threat posed by the norm-violating behaviour (i.e. violating the norms of one's ingroup) can therefore fulfill the conditions for a social identity to become salient (Ellemers et al., 2002; Hogg & Smith, 2007; Tajfel & Turner, 1979; Wenzel & Woodyatt, 2025). By making a group identity salient through his editorial actions, Biddle acts as a *social identity leader* that mobilises the group to engage in shaming.

Shamers as Engaged Followers of the Leader

There has been a long tradition in social psychology to examine how leader influences people's behaviour. Most interestingly, much of this literature has emerged in the context of examining and reanalysing Milgram's research that examined how people can be influenced to punish others. In the 1960s, Stanley Milgram set out to examine how authority could influence individuals to act against their own moral codes (Milgram, 1963). The experiments aimed to investigate whether people would obey an authority (e.g., the experimenter) even when they were asked to engage in harmful, inhumane actions (e.g., punishing a supposed learner with electric shocks) (Milgram, 1974). It was found that over 60% of the participants proceeded to the end of the study by administering a lethal level of electric shock to the learner. Milgram concluded that, even when the punishing behaviour itself would clearly cause severe and negative consequences, people still engage in such punishments due to their obedience to authority.

However, some researchers have argued that participants' engagement in the classic Milgram experiment should be understood as an outcome of social identity-based leadership (e.g., Haslam et al., 2015). In the reconceptualisation of Milgram's experiment, Reicher and colleagues (2012) found that people who *identified* more with the experimenter were more

willing to continue the experiment by punishing the learner. In other words, it was people's identification with the leader that made them willing to enact the values expressed by the leader. Indeed, it was suggested that in Milgram's study, the experimenter became a leader who represented a scientific enterprise and outlined the research objectives – clearly showing how the participation would improve society (Haslam et al., 2014; Haslam et al., 2015). Therefore, people engaged with the experimenter's order because they identified with the leader and the group-based goals that the leader represented, rather than being mindlessly obedient to authority.

Consistent with the results of the replication of the Milgram experiments, researchers proposed the *engaged followership model* to explain why people were willing to punish or hurt others (e.g., Haslam et al., 2015). Built on the social identity approach to understand leadership and social influence, the model suggests that when people are asked to perform unpleasant tasks by other individuals (such as a leader), people's willingness to engage in such tasks can be shaped by their social identification with such individuals, as well as by the belief that such individuals represent the shared goals and values. Since online shaming can be understood as a form of group-based behaviour with the goal of punishing the perceived wrongdoer (Chapter 2), the engaged followership model may be particularly useful in understanding why people engage in online shaming. Specifically, we suggest that whether people can be effectively mobilised by the leader whom they identify with could be contingent on 1) whether there is a punitive norm that specifies the appropriate response in the specific situation (i.e., engaging in online shaming) and 2) whether there is a noble goal that specifies the necessity to shame.

Leaders Create Social Influence via Group Norms and Goals

When people self-categorise as members of a group, their attitudes and behaviours tend to be influenced by the ingroup, especially through the group norms (Abrams et al., 1990; Hogg

et al., 1990; Turner, 1991). Group norms can be defined as shared expectations within a group that guide how people should feel, think, and behave (e.g., Postmes et al., 2000). People who identify more with a group tend to be more likely to conform to the group norms (Terry & Hogg, 1996; White et al., 2009). Furthermore, group norms are often not static and can be inferred from interactions with other ingroup members (such as emerging through ingroup communications; Postmes et al., 2000). In particular, group leaders can be influential in mobilising the group members' behaviour through demonstrating the norm (Reicher et al., 2005; Reicher et al., 2012; van Knippenberg & Hogg, 2003). This can be achieved by acting it out or verbally expressing their view about how the group members should behave (Haslam et al., 2010; Hogg & Giles, 2012; Hogg & Reid, 2006). Consistently, in the context of online shaming, we suggest that a leader presenting a punishing norm (via shaming expressions) with whom one identifies would be influential in mobilising people's behaviour. In the example of Justine Sacco, the leader (Sam Biddle)'s behaviour of shaming her for being a norm-violator (i.e., someone who made a racist remark reflecting white privilege) might have served as an exemplar behaviour, which motivated the other likeminded individuals (i.e., the ingroup members) to engage in shaming.

However, if the behaviour guided by the leader seems socially undesirable, demonstrating a norm alone might not be sufficient for an effective mobilisation. This argument can be supported by the reinterpretation of the classic Milgram experiments. Through reanalysing the archives of the Milgram experiments, researchers found that, instead of lack of distress, people actually felt good about the research aim and for being able to participate in the research (Haslam et al., 2015). Therefore, rather than unwillingly obeying the leader's orders, people were motivated to work towards the noble collective goal set by the leader (i.e., the scientific goal), which provided a rationale for the punishing behaviour (Reicher et al., 2012). Indeed, while

online shaming is often perceived as undesirable for its negative consequences, the presence of a noble goal can be found across different incidents of online shaming (Ronson, 2016). In previous chapters (Chapter 2 and 3), we also found that online shaming can entail the goal of deterring the perceived wrongdoing and/or creating change, both of which can be interpreted as noble by people who engage in shaming. Consistent with the engaged followership model explaining why people punish, in this research we wanted to test whether people can be mobilised to engage in online shaming behaviours. Especially, whether people can be mobilised by a leader they identify with, who sets a punishing norm alongside a noble group goal.

Overview

Across the two studies to be presented in this chapter, we examined whether online shaming as a group-based punishment can be mobilised by a leader through a shared social identity. We used two paradigms where the group goal was to elicit certain types of prosocial behaviour, in Study 3 the goal was reducing anti-social behaviour (somewhat ironically given the methods used), and in Study 4 (performed during the early emergence of COVID-19) was encouraging health compliance to COVID health responses. In Study 3, we used a paradigm developed from the original Milgram study to test the hypotheses suggested by the social identity theorists' reinterpretation of Milgram's study findings (Haslam et al., 2015; Reicher et al., 2012). Namely, whether a mobilising identity leader (i.e., who belongs to an ingroup and becomes a source of social identification, and who presents the norm of punitive behaviour as well as a noble goal to support the punitive behaviour) would be more likely to motivate others to engage in online shaming, compared to when the person involved was not a mobilising identity leader (i.e., who lacks a shared identity and does not provide the punitive norm or noble goal). In Study 4, to further investigate the leader's mobilisation, we disentangled the manipulations used in

Study 3. Specifically, we examined how people's identification with the leader interacts with the leader's influence (via the presence of norm and goal) in predicting people's online shaming engagement. Ethical approval was obtained from the Flinders Human Research Ethics Committee (8534) for Studies 3 and 4.

Study 3

In Study 3, participants were asked to train an Artificial Intelligence (AI) system that responds to people who post hostile comments, in order to reduce future online hostility. Participants were asked to respond to four AI training trials, with the options to select a pre-made shaming comment and contribute an optional shaming comment for each of the trials. Over the trials, hostility increased in the pre-made shaming comments, mimicking the increasing punishments in Milgram's study. To test whether people's online shaming engagement can be mobilised by an identity-based leader, participants were randomly allocated into one of the two conditions: 1) the mobilising identity condition, where participants were introduced to a leader who belongs to the ingroup and presents both a punitive norm and a noble goal and 2) a non-mobilising condition, where participants were introduced to a leader who belongs to an outgroup and lacks both the punitive norm and a noble goal.

Hypotheses

We hypothesised that although participants would become less likely to shame over the trials in general, when compared to the non-mobilising condition, participants in the mobilising identity condition would be more likely to engage in each of the trials by selecting a shaming comment. Their selected shaming comments would also be more hostile than those in the non-

mobilising condition[3]. And lastly, the mobilising condition would be more likely to contribute an optional shaming comment for each of the trials than non-mobilising condition.

Method

Participants and Design

Two hundred participants were recruited via Amazon Mechanical Turk (MTurk) to participate in a short online study with the inclusion criteria of being a US citizen, Caucasian and a frequent user of social media. Participants who answered at least three of six checks (including three comprehension checks and three attention checks) incorrectly were excluded from the analyses. A total of 174 participants were included in the study. Participants' ages ranged from 20 to 72 (M = 38.49, SD = 12.11), 85.1% of them used social media several times a day, 12.1% used once a day, 1.7% used once every two to three days, and 1.1% used once a week. Participants were randomly assigned to the mobilising identity condition (N = 80; 51.2% Female, 1.3% non-binary) or non-mobilising condition (N = 94; 54.3% Female, 1.1% non-binary). With this sample size, we conducted sensitivity analyses using the program G*Power 3.1 (Faul et al., 2009). An ANOVA analysis with one predictor, power of 0.80 and alpha is 0.05 would have a sensitivity of detecting a minimal explained variance effect size of f = 0.21, a small to medium effect size (Cohen, 1988). Assuming the probability of someone in the non-mobilising condition engage in shaming is 0.15, a two-tailed logistic regression with a binominal predictor and a power of 0.80

[3] Although the comments selected for each of the trials were at a similar level of hostility, some were rated as slightly more hostile than others (see pilot results presented in Table C3). Therefore, we computed weighted shaming choices as a set of dependent variables, separate from the categorical dependent variables, taking the mean hostility scores into account.

at the significance level of 0.05 would have a sensitivity of detecting an odds ratio of 2.82, a small to medium effect (Ferguson, 2009).

Procedure

Potential participants from MTurk were asked to specify how often they used social media. Participants who used social media once a week or more were randomly assigned to a mobilising identity or a non-mobilising condition, and were shown different versions of information sheet accordingly. Participants in the mobilising identity condition were told that this AI training program was organised by a Non-Governmental Organisation in the US, whereas participants in the non-mobilising condition were told that the organiser of the program was a Chinese IT company. After providing informed consent, participants read a message from the CEO. The ingroup CEO specified a noble goal of building a hostility-free online environment, and the outgroup CEO specified a less noble goal of developing a top-selling AI product. Moreover, only the ingroup CEO presented an explicit group norm in the message that responding with disapproving comments to people who engage in online hostility can be an effective way of reducing online hostility. After answering comprehension check questions that asked about the content of the CEO's message, participants were told that the AI training would start on the next page. After the trials, participants responded to questions about the group norm, identification with the leader, and the goal[4]. Finally, they were asked to provide demographic

[4] Participants were then asked about how deserving they perceive the commenter in the last trial to be shamed because of the replied hostile comment to the post, as well as other variables such as their sense of contribution to the issue of preventing online hostility (see Appendix C for the measures and results of the other variables).

information, including age, gender, citizenship, cultural or ethnic background, and the social media platforms they use.

Stimulus Materials

We created 4 social media posts of that present online hostility as the norm violation. Each social media post included 1) a person who posted an image with a few sentences about the image, called the original post and 2) a person replied to the original post with a hostile comment (see Figure 12 for example). The person's reply constituted an example of online hostility, to which the participants could choose to respond with online shaming. The social media posts were pre-tested among the postgraduate students in psychology and friends of the principal researcher. In the pilot study, participants ($N = 27$[5]) were asked for each of the posts: "How hostile is the comment replying to the original post?" on a 5-point scale where 1 = *Not hostile at all*, 5 = *Extremely hostile*. Four posts, in which the hostile replies were rated as moderately hostile to very hostile, were selected for Study 3.

Participants in study 3 were told that the organiser behind the study aims to develop an AI program that reduces online hostility, and their task was to train the AI program. Participants were then asked to engage in four trials of AI training and were shown the pre-selected social media posts. For each of the posts, participants were presented with 3 shaming options and an opt-out option (i.e., "AI should not respond")[6], and they were instructed to choose one of the four

[5] We failed to gather 6 participants' rating about the hostility of comments appearing in the social media posts due to an issue of operating Qualtrics.

[6] For each of the trials, the three options of shaming comments were presented in a random order, whereas the opt-out option were always presented as the fourth option.

options as a response to the person who replied with a hostile comment. The shaming options are comprised of shaming comments that were pre-tested (Table C2). To mimic the incremental punishments in Milgram's paradigm over trials, the shaming options for the participants to choose from were presented in the order from the least hostile comments, slightly hostile comments, hostile comments, to extremely hostile comments. Although the social media posts were presented in different orders[7], the shaming comments were always presented in the order from the least hostile to extremely hostile over trials. The social media posts and shaming comments altogether comprised the stimuli used in the first study. Appendix C provides detailed information on the pre-testing of the study stimuli.

[7] The social media posts were either in the normal order of posts 1, 2, 3 and 4, or in the reversed order of posts 4, 3, 2 and 1.

Figure 12

Example Social Media Post Used in Study 3

Note. This example social media post is comprised of a neutral, original post and a hostile reply. The hostile reply represents a norm violation of online hostility, in which the participants can choose to respond with online shaming.

Measures

Participants responded on scales anchored 1 (*strongly disagree*) to 7 (*strongly agree*) unless otherwise specified.

Manipulations

Agreement With the Norm. Two items were used to measure agreement with the norm: "I find it important to educate people who engage in online hostility." and "I think responding with disapproving comments to people is an effective way of stopping their hostile behavior.", $r = .30$, $p < .001$. Participants' scores were averaged to create a single scale, with higher scores reflecting higher agreement with the leader's norm.

Identification With the Leader. Two items, adapted from Haslam et al. (2014), were used to measure the identification with the leader: "I identify with the CEO who initiated the AI training program." and "I feel positive about the organizers behind the AI training program.", $r = .71$, $p < .001$. Participants' scores were averaged to create a single scale, with higher scores reflecting higher identification with the leader.

Identification With the Goal. Two items, adapted from Haslam et al. (2014), were used to measure identification with the goal: "I think the goal of the organization/company is important." And "I want to help the organization/company to achieve its goal.", $r = .65$, $p < .001$. Participants' scores were averaged to create a single scale, with higher scores reflecting higher identification with the goal.

Shaming Responses

Participants' shaming responses were evaluated, including their chosen shaming comments and the contributed shaming comments, with higher scores reflecting a higher level of

engagement in online shaming behaviour. For all shaming trials, the primary and the secondary researchers coded each of the additional comments contributed by the participants.

Shaming Choice by Trial. Four different scales were created for the four trials of AI training (i.e., Trial 1 shaming choice, Trial 2 shaming choice, Trial 3 shaming choice, and Trial 4 shaming choice). We recoded the possible choices participants could have selected in each of the trials: choosing one of the three shaming comments they were presented with or the opt-out option "AI should not respond" (0 = *Choose to not respond*, 1 = *Choose to respond*).

Weighted Shaming Choice by Trial. Although the comments selected for each of the trials were at a similar level of hostility, some were rated as slightly more hostile than others (results presented in Table C3). Therefore, four different weighted scales were created for the four trials of AI training (i.e., Weighted trial 1, Weighted trial 2, Weighted trial 3, and Weighted trial 4). The scores were calculated as the product term of the Shaming Choice by Trial (0 = *Choose to not respond*, 1 = *Choose to respond*) × mean hostility of the chosen comment calculated from the pilot study (1 = *Not hostile at all*, 5 = *Extremely hostile*). For example, a participant who chose the comment "Your comment shows an underdeveloped level of maturity." in the second trial obtained a score of 2.61 on Weighted trial 2. The score 2.61 was the mean hostility of the chosen comment rated by those who participated in the pilot study (see Appendix C for further details). And a participant who chose "AI should not respond" in the second trial scored 0 on Weighted trial 2.

Contributed Comments by Trial. In addition to choosing from provided shaming comments for each of the shaming trials, the participants were also offered an opportunity to provide a comment to the person who replied with a hostile comment, and were told that the comment would be added in the future AI training. While being unaware about which condition

each participant belonged to, the primary and secondary researchers independently coded the provided comments into the following categories: 0 = *Response absent*, 1 = *Showing an agreement with the shamer*, 2 = *Providing an evading response*, 3 = *Showing a disagreement with the shamer*, 4 = *Shaming/calling out the shamer's behaviour*, 5 = *Shaming the shamer*, 6 = *Unclear*. Disagreement between the two coders were resolved through discussion. To prepare for subsequent analyses, we further collapsed categories with low frequencies: The categories 0, 1, 2 and 6 were collapsed into the new category 1 = *Non-punitive/No responses*, category 3 was recoded into 2 = *Disagreeing response*, and the categories 4 and 5 were collapsed into the new category 3 = *Shaming response*. Among these categories, we were most interested in the shaming responses, as they indicate a further engagement with the leader. Examples of shaming responses are "Your comment was rude, mean, judgemental, and unacceptable" and "You're a sick person if you imagine things like that when seeing a cute photo of a dog".

Results

The descriptive statistics and inter-correlations between variables are shown in Table 2. Preliminary evidence from the study descriptive statistics suggested that participants who were in the mobilising identity condition reported higher identification with the leader than those who were in the non-mobilising condition. Also, higher identification with the leader was correlated with higher hostility of the chosen shaming comment in the third trial as well as for the average weighted shaming choice across trials.

Table 2

Correlations with Means and Standard Deviations by Condition

Variable	Mobilising[a] *M*(*SD*)	Non-Mobilising[b] *M*(*SD*)	1	2	3	4	5	6	7
1. Identification with leader	4.95(1.22)	4.38(1.28)							
2. Agreement with norm	4.50(1.31)	4.05(1.26)	.55**						
3. Identification with goal	5.49(1.08)	4.97(1.24)	.80**	.47**					
4. Weighted trial 1	1.12(0.76)	0.78(0.82)	.13	.26**	.12				
5. Weighted trial 2	1.89(1.17)	1.72(1.24)	.12	.16*	.12	.20**			
6. Weighted trial 3	1.23(1.68)	1.23(1.68)	.24**	.32**	.23**	.17*	.36**		
7. Weighted trial 4	0.57(1.44)	0.48(1.32)	.07	.14	.11	-.09	.14	.30**	
8. Average weighted choice	1.20(0.80)	1.05(0.84)	.23**	.35**	.24**	.37**	.66**	.81**	.60**

Note. $N = 174$. Standard deviations are presented in parentheses.

[a] n = 80. [b] n = 94.

*$p < 0.05$ level, 2-tailed. **$p < 0.01$ level, 2-tailed.

Manipulation Checks

A one-way MANOVA was used to examine whether the manipulations were successful. It showed that conditions had a significant effect on the combined dependent variables (i.e., identification with the leader, agreement with the norm, and identification with the goal), Pillai's trace = .06, $F(3, 170) = 3.41$, $p < .05$. Univariate ANOVAs showed there were significant differences between conditions for identification with the leader, $F(1, 172) = 8.87$, $p < .01$, $\eta^2_p = .05$, agreement with the norm, $F(1, 172) = 8.63$, $p < .05$, $\eta^2_p = .03$, and identification with the goal, $F(1, 172) = 8.53$, $p < .01$, $\eta^2_p = .05$. Participants in the mobilising identity condition reported higher identification with the leader, a higher agreement with the leader's norm, and a higher identification with the leader's goal than those in the non-mobilising condition. Thus, the manipulations were successful.

Comparison between Conditions

Shaming Choice by Trial. The percentages of participants who selected a shaming choice in each trial by condition are presented in Table 3. In general, there was decremental engagement in the shaming behaviours over trials in both conditions (except for the second trial). A series of logistic regression was performed to test the effect of condition on the likelihood that participants select a shaming comment for each trial, which are presented in Table C3 of Appendix C. Order was included as a covariate in the logistic regressions, as differences in the content of the social media posts might influence participants' online engagement in a given trial. Condition was found to predict people's behaviour of selecting a shaming comment only for the first trial[8], B = -

[8] The covariate, Order, was also significant in predicting the Trial 1 shaming choice.

0.88, $SE = 0.33$, $p < .01$. The odds ratio of 0.42 for condition indicates that the odds of choosing a shaming comment are 0.42 less when participants were in the non-mobilising condition than in the mobilising identity condition. That is, the odds are decreased by 58%. However, such a pattern was not found in the other trials. Thus, the hypothesis that, relative to the non-mobilising condition, participants in the mobilising identity condition would be more likely to engage in shaming was supported for the first trial of AI training, but not for the second to fourth trials.

Weighted Shaming Choice by Trial. A set of one-way ANCOVAs was conducted to compare the weighted shaming choices between the mobilising identity and non-mobilising conditions, with order as the covariate. A significant difference between conditions was only found for the first trial, $F(1, 170) = 8.26$, $p < .01$, $\eta^2_p = .04$, but not for the second trial, $F(1, 170) = 0.88$, $p = .35$, $\eta^2_p = .01$, the third trial, $F(1, 170) < 0.01$, $p = .99$, $\eta^2_p < .01$, the fourth trial, $F(1, 170) = 0.20$, $p = .66$, $\eta^2_p < .01$, nor the average weighted shaming choice, $F(1, 170) = 1.40$, $p = .24$, $\eta^2_p = .01$. Thus, the hypothesis that, relative to the non-mobilising condition, participants in the mobilising identity condition would select shaming comments that are more hostile was supported for the first trial of AI training, but not for the second to four trials or the average weighted shaming choice.

Table 3

Percentages of Participants Selected a Shaming Choice by Condition and Trial

Trial	Mobilising Identity[a]	Non-Mobilising[b]
Trial 1 shaming choice	68.8%	47.9%
Trial 2 shaming choice	72.5%	66.0%
Trial 3 shaming choice	35.0%	35.1%
Trial 4 shaming choice	13.8%	11.7%

Note. $N = 174$.

[a] n = 80. [b] n = 94.

Contributed Comments by Trial. For the contributed comments, the percentages of participants who responded differently by condition were presented in Table 4. It can be seen that for each of the trials, whether participants were in the mobilising identity condition or in the non-mobilising condition, a higher percentage of them left a non-punitive or no response, and a lower percentage of participants left a disagreeing response, compared to those who left a shaming response.

A series of multinomial logistic regression was performed to create models of the relationship between the condition and the three types of contributed comment responses (non-punitive/no response, disagreeing response, and shaming response) for each of the four trials, with order included as the covariate. The results are reported in Tables C4–C6 of Appendix C. For each of the trials, a test of the model with all predictors against a constant-only model was conducted. And it was found to be significant only for the Trial 3 contributed comments, χ^2 (6) = 15.27, $p < .05$, indicating that the predictors, as a set, significantly predicted the different types of comment responses for Trial 3. As shown in Table 5, the main effect of condition and the interaction term between condition and order significantly predicted the disagreeing response (compared to non-punitive/no response) but not the shaming response (compared to non-punitive/no response). The main effect of condition indicates that compared to participants in the non-mobilising condition, those who were in the mobilising identity condition were more likely to leave a disagreeing response than a non-punitive or no response in Trial 3. An examination of the marginal effects on the probability for each type of response further revealed that when

participants were presented trials in the normal order than in the reversed order[9], those who were introduced to a mobilising leader were less likely to leave a disagreeing response (b = -0.25, SE = 0.09, z = -2.95, $p < 0.01$), and were more likely to leave a non-punitive or no response (b = 0.34, SE = 0.11, z = 3.26, $p < 0.01$). However, when participants were introduced to a non-mobilising leader, the effect of order was not significant for disagreeing response (b = -0.01, SE = 0.06, z = -0.22, p = 0.83), nor for non-punitive or no response (b = -0.01, SE = 0.10, z = -0.10, p = 0.92, 95% CI = [-0.21, 0.19].

Overall, we did not find support for the hypothesis that people in the mobilising identity condition would be more likely to leave additional shaming comments than those who were in the non-mobilising condition. However, we found that compared to people who were in the non-mobilising condition, those who were in the mobilising identity condition were more likely to show a disagreement with the commenter in the third trial, though such an effect differed based the social media posts.

[9] Participants who were in the normal order condition viewed the social media post "Dog in snow", and those who were in the reversed order condition viewed the social media post "Homelessness" (see Appendix C).

Table 4

Percentages of Participants' Responses by Condition and Type of Contributed Comments

	Mobilising Identity[a]				Non-Mobilising[b]			
	Trial 1	Trial 2	Trial 3	Trial 4	Trial 1	Trial 2	Trial 3	Trial 4
Non-punitive/No response	38.75%	57.50%	56.25%	48.75%	48.94%	60.64%	62.77%	55.32%
Disagreeing Response	23.75%	11.25%	18.75%	16.25%	15.96%	11.70%	9.57%	21.28%
Shaming response	37.50%	31.25%	25.00%	35.00%	35.11%	27.66%	27.66%	23.40%

Note. $N = 174$.

[a] n = 80. [b] n = 94.

Table 5

Multinomial Logistic Coefficients for Predicting Trial 3 Contributed Comments

Predictor	Trial 3		
	B (*SE*)	Wald	OR [95% CI]
	Disagreeing vs. Non-punitive/No responses		
Condition[a]	-1.67 (0.62)	7.24*	0.19** [0.06, 0.64]
Order[b]	-2.18 (0.72)	9.12**	0.11** [0.03, 0.46]
Condition × Order	2.06 (1.02)	4.08*	7.84* [1.06, 57.8]
Constant	-0.15 (0.39)	0.15	NA
	Online shaming vs. Non-punitive/No responses		
Condition[a]	-0.63 (0.52)	1.44	0.53 [0.19, 1.48]
Order[b]	-1.00 (0.55)	3.24	0.37 [0.13, 1.09]
Condition × Order	1.10 (0.73)	2.28	3.00 [0.72, 12.40]
Constant	-0.24 (0.40)	0.36	NA

Note. $N = 174$. OR stands for odds ratio.

[a] Condition was coded as 1 = mobilising identity (the reference group; n = 80), 2 = non-mobilising (n = 94). [b] Order was coded as 1 = normal (the reference group; n = 86), 2 = reversed (n = 88).

* $p < .05$. ** $p < .01$

Discussion

The manipulations of people's identification with the leader and agreement with the leader's norm and goal were successful. However, the proposed hypotheses were generally not supported, except for the first trial when the shaming options were only mildly hostile. It was found that relative to the non-mobilising condition, participants in the mobilising identity condition were more likely to engage in shaming only in the first trial by choosing a shaming option which was more hostile. Particularly, we saw an increased disengagement in shaming over trials: As the shaming options became more and more hostile, participants became more and more disengaged (whether they were introduced to a mobilising or a non-mobilising leader) from choosing a shaming comment. When the participants were shown very hostile comments (as in the last trial), only about 10% of the participants engaged in choosing a shaming comment. This is consistent with Milgram (1974)'s findings, in which more people started to feel reluctant to exert the punishments when they became more severe.

Similarly, the proposed hypothesis that participants in the mobilising identity condition would be more likely to engage in shaming than participants in the non-mobilising condition was not supported for the optional comments that participants contributed. It was found in the third trial that for participants who were shown the social media posts in the reversed order, people in the mobilising identity condition were more likely to show disagreement (with the person who expressed hostility) than those who were in the non-mobilising condition. However, people who were in the mobilising identity condition did not engage in shaming more than those in the non-mobilising condition. Again, we observed that participants were generally not prompting the system to engage in online shaming. For each of the trials, the majority of participants did not contribute a shaming comment; instead, they preferred not to leave a comment or to leave a non-

punitive comment. However, since we only had two conditions (mobilising vs. non-mobilising), we were unable to separate the influences of the presence of leader's norm and goal and leader's group membership on people's shaming engagement. This issue is addressed in Study 4 by further disentangling the manipulations.

In sum, it was found that people generally disengaged with the leader when it comes to online shaming, as shown from their disengagement in both choosing and contributing shaming comments. Due to this disengagement, it is difficult to examine whether people can be effectively mobilised by the leader whom they identify with to engage in shaming. To address this issue, in Study 4, we used a behavioural measure of online shaming that does not require people to directly choose or contribute a shaming comment. Rather, people would be asked about their behavioural intention to show support for some social media posts (via liking, sharing, and commenting) that involves shaming content. In an experimental setting, people might be more likely to indicate their behavioural intention to engage in online shaming than to exert harsh punishments themselves. However, it is also possible that people continue to show reluctance to shame. We suggest the reluctance to shame could be reflected in people's disidentification with the ingroup leader (Becker & Tausch, 2013; Chien, 2024), who employs shaming as a punishment. Therefore, in Study 4, we measured people's identification with the leader either before or after they are exposed to shaming, to examine whether there is a difference in their identification with the leader. Lastly, two attitudinal measures (i.e., perceived appropriateness and perceived effectiveness) were included to further examine the mobilising effects of an identity-based leader in shaping people's view about online shaming.

Study 4

Study 4 was conducted early during COVID-19 pandemic (i.e., in 2020) to examine support for harsh and punitive shaming methods being used to elicit health message compliance. In Study 4, we further disentangled the leader's mobilisation, by examining how people's identification with the leader interact with the leader's influence via the presence of norm and goal in predicting people's attitudes of online shaming and online shaming engagement (in a milder form). Participants were randomly assigned to one of the eight conditions (2×2×2 design), based on whether the participants were introduced to an ingroup or an outgroup leader, whether the leader's noble goal and norm were presented, and the number of shaming comments (one comment vs. five comments). The dependent variables are: 1) behavioural measures, including people's intention to like, share, and willingness to comment on posts that contain shaming content and 2) people's attitudes towards the use of online shaming, including perceived appropriateness and perceived effectiveness of online shaming. The study further examined whether people would still be reluctant to engage in shaming, which can be reflected from their (dis)identification with the leader. That is, people are asked about their identification with the leader, either before or after viewing shaming punishments. We then compare their identification with the leader and see if there would be a difference.

Different from Study 3, in Study 4 we measured people's *behavioural intention* to support a campaign that utilises online shaming, in the form of "liking", "sharing" or willingness to comment on a post (e.g., Barron et al., 2023). We suggest that this new behavioural measure is appropriate because online shaming can be deemed as socially undesirable. Participants might be unwilling to demonstrate shaming behaviour that takes a stronger form (e.g., posting shaming content themselves), but be willing to engage in milder forms of behaviour that are more

indirect, such as showing one's support for others' shaming behaviour. Additionally, we examined whether people's perceived appropriateness and effectiveness of online shaming would differ when the shaming goes viral as opposed to a non-viral condition. It was found that when more people expressed their outrage online (which often involves shaming), people believed it was normative to express condemnation but also felt more sympathy towards the perceived wrongdoer (Sawaoka & Monin, 2018, 2020). We manipulated the number of shaming comments, to explore how an aggregated (viral) punishment influences people's attitudes and intention to support online shaming.

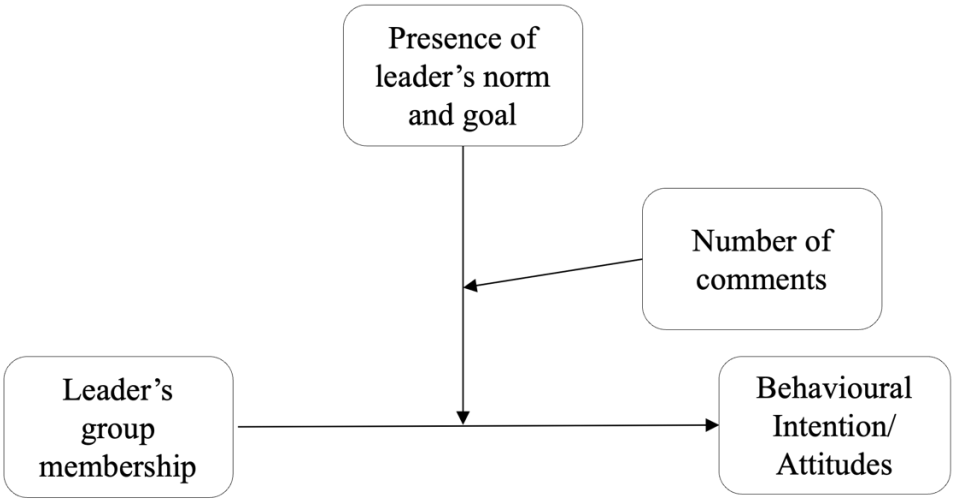
Hypotheses

Consistent with Study 3, the first hypothesis (H1) posited that people would show a greater intention to support online shaming and find it more acceptable and effective when presented with an ingroup leader (i.e., operationalised as a leader same nationality) rather than an outgroup leader (operationalised as a leader from a different nationality). We also examined whether this predicted relationship would be further moderated by the presence of leader's norm and noble goal combined. Specifically, we expected that the relationship between leader's group membership and people's behaviour intention (H2a), as well as the relationship between leader's group membership and their attitudes about online shaming (H2b), would be stronger when people are presented with a punitive norm together with a noble goal, than when those who are not presented with such information. Additionally, we explored if there is a three-way interaction (see Figure 13 below). Specifically, the two-way interaction between leader's group membership and the presence of leader's norm and goal was expected to be further qualified by the number of shaming comments (H3). Finally, in line with the suggestion that people might be reluctant to engage in online shaming, we predicted that people's identification with the leader would be

lower if it is measured after they have viewed the shaming posts, than when it is measured before viewing the posts. We expected this effect to be stronger for an ingroup leader than for an outgroup leader, due to a disidentification with the ingroup leader. Therefore, we hypothesised a two-way interaction between the order of the identification with the leader (either measured before or after) and leader's group membership in predicting people's identification with the leader (H4).

Figure 13

Conceptual Model of the Hypothesised Moderated Moderation



Method

Participants and Design

Data was collected from May 4, 2020, to May 9, 2020 (UTC). Four hundred and forty-nine participants (51.5% were female, 0.7% were non-binary) were recruited via MTurk to participate in a short online study with the inclusion criteria of being a US citizen, Caucasian,

and a frequent user of social media. Participants who answered at least two of the five checks (including three comprehension checks and two attention checks) incorrectly were excluded from the analyses. A total of 406 participants were included in the study. Participants' ages ranged from 18 to 77 (M = 36.89, SD = 12.98), 85.5% of them used social media several times a day, 10.6% used once a day, 2.7% used once every two to three days, and 1.2% used once a week. Participants were randomly assigned to one of the eight conditions shown in Table 6. With this sample size, we conducted sensitivity analyses using the program G*Power 3.1 (Faul et al., 2009). A linear multiple regression with seven predictors, one covariate (controlling for the order of identification with leader questions), power of 0.80 and alpha is 0.05 would have a sensitivity of detecting a minimal explained variance effect of $f^2 = 0.04$, a small to medium effect size (Cohen, 1988). An ANOVA analysis with three predictors, power of 0.80 and alpha is 0.05 would have a sensitivity of detecting a minimal explained variance effect size of $f = 0.19$, a small to medium effect size (Cohen, 1988).

Stimulus Materials

Participants were told that they would evaluate some materials used in a public health campaign, that aims to address people's behaviour during COVID-19 pandemic. We created 3 social media posts that deployed online shaming to address the following behaviour respectively: stockpiling, attending a party, and sending a child with coughing symptoms to school (see Figure 14 for an example). Online shaming of these behaviours was prevalent since the early stage of the COVID-19 pandemic (Max, 2020; Tait, 2020). Each post was composed of a screenshot of a fabricated Facebook post made by someone who did not comply with the behavioural guidelines, either one or five shaming comment(s) towards the person who made the post, as well as a sentence "Think about what you SHOULD do to prevent coronavirus" to deliver the message of

the campaign. The shaming comments included in the posts were pre-tested in a pilot study. Details of the pilot study appear in Appendix C, and all shaming posts are reproduced in Appendix D.

Figure 14

Example Social Media Campaign Post Used in Study 4

Procedure

Similar to Study 3, potential participants on MTurk were first asked to specify how often they used social media. Participants who used social media once a week or more were randomly assigned to either an ingroup or an outgroup condition. They were shown different versions of information sheets about the campaign, as either organised by an American leader from an American organisation based in the US, or by a Chinese leader from a Chinese organisation based in China. After providing informed consent, participants were shown a message from the CEO. There were four different versions of the message, based on the leader's group membership (ingroup vs. outgroup) and whether the leader's norm, combined with a noble goal, was presented. Participants who were presented with the leader's norm and goal were told that it was right and moral to follow the COVID-19 behaviour guidelines, as it would help to flatten the curve of the spread of COVID-19 and protect the community (i.e., emphasised the nobleness of the goal). They were also told that responding with disapproving comments towards those who did not follow the guidelines can be an effective way of educating people and reinforce them to follow the guidelines (i.e., a norm to punish the non-compliers). For participants who were allocated to the condition in which the leader's norm and goal were absent, the information about norm was not presented, nor was the noble goal emphasised.

The manipulation check questions about the leader's norm were presented after viewing the CEO's message. This was followed by a measure of identification with the leader (for only half of the participants). Thereafter, participants were randomly assigned to either of the two conditions: one shaming comment and five shaming comments. Following viewing each post, participants were asked to indicate their intention to support the campaign (via liking, sharing, commenting), perceived effectiveness of the posts, and their perceived appropriateness of the last

(or the only) shaming comment presented in the post. Next, we measured participants' identification with the leader (for only half of the participants who were not asked about their identification earlier). Then, we measured all participants' identification with the goal, perceived nobleness of the goal, as well as other variables such as their feelings regarding the pandemic and their experience of using social media during the pandemic. Finally, participants were asked to provide demographic information, and indicate their view about the campaign and what they thought this research was about. Some comprehension check questions were included in the survey. All participants performed well on the comprehension checks. Details of the study manipulation and a complete list of variables are provided in Appendix D.

Measures

Participants responded on scales anchored 1 (*strongly disagree*) to 7 (*strongly agree*) unless otherwise specified.

Manipulations

Identification With the Leader. We used the same items as in Study 3 (except the "AI training program" was replaced by "public health campaign"), $r = .81$, $p < .001$. Scores were averaged to create a single scale, with higher scores reflecting higher identification with the leader.

Identification With the Goal. We used the same items as in Study 3 (except the "AI training program" was replaced by "public health campaign"), $r = .67$, $p < .001$. Scores were averaged to create a single scale, with higher scores reflecting higher identification with the goal.

Perceived Norm. Participants were asked to indicate whether they perceived a norm represented by the leader: "For the purpose of this campaign, I know what we should do with

people who failed to comply with the guidelines", with a response of 0 = *No*, 1 = *Yes*. For those who selected "Yes" for this question, they were asked an additional open-ended question to recall the punishing norm: "What should we do with people who failed to comply with the guidelines?". The primary researcher coded responses to this question to determine whether participants accurately recalled the norm's content. Coding details are provided in Appendix D.

Dependent Variables

In contrast to Study 3, where online behaviour measures were analysed by trial due to the varying levels of hostility in the shaming comments presented to participants, the shaming comments in the current study were consistently at a similar level of hostility. Therefore, the dependent variables presented below were collapsed across posts and averaged into a single scale.

Behavioural Intention to Support. For each of the posts, three items were used to measure participants' behavioural intention to support the post on social media: 'I would "like"/share/comment under the post[10] on social media to support the campaign.': The intention to show support via interacting with the post that addresses stockpiling behaviour, α = .93, partying, α = .90, sending kids to school while coughing, α = .91, and overall, α = 0.94. Higher scores reflected a greater intention to support the campaign.

Perceived Appropriateness. For each of the posts, three items were used to measure the perceived appropriateness of the last shaming comment: "I find [commenter's name]'s response

[10] In the original survey questions, each of the social media posts was referred to as a "poster", to distinguish from the fabricated Facebook posts.

acceptable", "I think [commenter's name]'s response is appropriate", and "It is unacceptable that [commenter's name] responded in that way" (reverse coded). The perceived appropriateness of shaming comment involving stockpiling, $\alpha = .86$, attending a party, $\alpha = .89$, and sending a child with coughing symptoms to school, $\alpha = .91$, and overall, $\alpha = 0.90$. Higher scores reflected a greater perceived appropriateness of the shaming comment.

Perceived Effectiveness. For each of the posts, two items were used to measure the perceived effectiveness of each shaming post used in the campaign: “I believe this post will be effective in reducing people’s … behavior.” And “I believe this post will be useful in educating people to not engage in …”: The targeted behaviours were stockpiling, $r = .83$, $p < .001$, partying, $r = .85$, $p < .001$, and parents sending a child with coughing symptoms to school, $r = .91$, $p < .001$, and overall, $\alpha = 0.90$. Higher scores reflected a greater perceived effectiveness of the shaming post.

Perceived Nobleness. Two items were used to measure the perceived nobleness of the goal: "I think the goal of the public health campaign is noble." and "I believe the goal of the public health campaign is moral.", $r = .78$, $p < .001$. Participants’ scores were averaged to create a single scale, with higher scores reflecting higher perceived nobleness of the goal.

Results

The descriptive statistics are shown in Table 6, and the inter-correlations between variables are shown in Table 7. It can be observed from the descriptive statistics that in general, people showed low intentions to provide a behavioural support for the campaign that used online shaming, and perceived online shaming as inappropriate and ineffective. The correlations showed that higher identification with the leader was associated with a greater intention to

support the posts, a higher perceived appropriateness of the shaming comments, and a higher perceived effectiveness of the shaming posts.

Table 6

Mean and Standard Deviations for Study Variables (Study 4)

Variable	Leader's norm and goal presented				Leader's norm and goal absent			
	Ingroup leader		Outgroup leader		Ingroup leader		Outgroup leader	
No. of Comment(s)	One (n = 49)	Five (n = 48)	One (n = 48)	Five (n = 58)	One (n = 55)	Five (n = 44)	One (n = 53)	Five (n = 51)
Identification with leader								
Measured before	4.64(1.56)	4.54(1.95)	4.03(1.66)	4.33(1.78)	5.22(1.02)	5.64(0.88)	4.99(1.09)	5.03(1.07)
Measured after	4.66(1.61)	4.16(1.63)	3.90(1.60)	3.92(1.95)	3.63(1.51)	3.45(1.91)	3.89(1.55)	3.88(1.93)
Identification with goal	5.44(1.38)	5.59(1.30)	5.61(1.29)	5.47(1.51)	5.53(1.13)	5.81(1.28)	5.80(0.92)	5.83(0.88)
Behavioural intention	2.88(1.33)	3.08(1.62)	2.57(1.48)	2.79(1.66)	2.40(1.14)	2.40(1.35)	2.20(1.11)	2.60(1.20)
Appropriateness	2.93(1.37)	2.50(1.35)	3.15(1.43)	2.69(1.34)	2.65(1.26)	2.86(1.39)	2.53(1.31)	2.29(1.24)
Effectiveness	3.04(1.34)	3.36(1.76)	2.81(1.34)	3.44(1.73)	2.70(1.25)	2.88(1.43)	2.50(1.30)	3.22(1.43)
Nobleness	5.35(1.43)	5.39(1.37)	5.44(1.42)	5.09(1.66)	5.67(1.14)	5.42(1.53)	5.69(1.02)	5.76(0.84)

Note. [a] The identification with the leader that was measured before the evaluation of the posts (n = 196). Standard deviations are presented in parentheses.

Table 7

Correlations for Study Variables

Variable	1	2	3	4	5
1. Identification with leader	-				
2. Identification with goal	.42**				
3. Behavioural intention	.46**	.28**			
4. Perceived appropriateness	.39**	.20**	.23**		
5. Perceived effectiveness	.44**	.30**	.26**	.79**	
6. Perceived nobleness of goal	.42**	.76**	.29**	.51**	.45**

Note. * indicates $p < .05$. ** indicates $p < .01$.

Manipulation Checks

A series of one-way ANOVAs was conducted to examine whether the manipulations were successful. For the manipulation of leader's group membership, since we measured people's identification with the leader either before or after, we conducted two separate ANOVAs. For participants who were asked about their identification with the leader before viewing the shaming posts, leader's group membership significantly predicted people's identification with the leader, $F(1, 194) = 3.96$, $p < .05$, $\eta^2_p = .02$, suggesting the manipulation for leader's group membership was successful. For participants who were asked about their identification with the leader after viewing the shaming posts, leader's group membership did not predict people's identification with the leader, $F(1, 208) = 0.28$, $p = .60$, $\eta^2_p < .01$.

We conducted a one-way MANOVA for the manipulation of the leader's goal. It showed that the presence of leader's norm and goal had a significant effect on the combined dependent

variables (i.e., identification with the goal and the perceived nobleness of the goal), Pillai's trace = .02, $F(2, 403) = 3.49$, $p < .05$. Univariate ANOVAs showed the presence of leader's norm and goal did not predict people's identification with the goal, $F(1, 404) = 2.96$, $p = .09$, $\eta^2_p = .01$, but predicted people's perceived nobleness of the goal, $F(1, 404) = 11.90$, $p < .01$, $\eta^2_p = .02$. However, it was found that people who were presented more information about the leader's norm and goal perceived the goal as less noble, compared to those who were presented with less information about the leader's norm and goal, suggesting the manipulation of goal was not successful. To examine if the manipulation of the leader's norm was successful, a series of logistic regressions was performed to examine the effect of the presence of leader's norm and goal on people's perception of the leader's norm. Detailed results are reported in Table D2 of Appendix D. It was found that a higher percentage of participants in the norm and goal presented condition understood the norm and recalled the content of the norm correctly than those who were in the norm and goal absent condition, suggesting the manipulation of leader's norm was successful. In sum, the manipulation of leader's norm, and the manipulation of leader's group membership based on those who were asked about their identification with the leader before viewing the shaming posts, were successful. However, the manipulation for the leader's goal was unsuccessful.

Hypothesised Moderated Moderation

PROCESS Model 3 (Hayes, 2017) was used to test the hypothesised moderated moderation between leader's group membership, the presence of leader's norm and goal, and the number of shaming comments in predicting people's behavioural intention, perceived appropriateness, and perceived effectiveness, respectively. Because the order of identification with the leader (either asked before or after viewing the shaming posts) was found to influence

participants' identification with the leader, we included it as a covariate. The unstandardised coefficients are shown in Table 8 below. It can be seen that the proposed three-way interaction was not significant for any of the outcome variables (behavioural intention, appropriateness, effectiveness), thus H3 was not supported. The proposed two-way interaction between the leader's group membership and the presence of leader's norm and goal was not significant for any of the outcome variables, thus H2a and H2b were not supported either. And lastly, the main effects of the leader's group membership predicting people's behavioural intention, perceived appropriateness, and perceived effectiveness were not significant. Therefore, H1 that people would have a greater intention to engage in shaming and perceived online shaming as more appropriate and effective when they were presented with an ingroup leader rather than an outgroup leader, was not supported. In sum, H1-3 were not supported.

Table 8

Regression Coefficients for Variables and Interactions Predicting Behavioural Intention to Support

Predictor	*b* (*SE*$_b$)	95% CI for *b*
Outcome variable: Behavioural intention, R^2 = .04, *MSE* = 1.89, *F*(8, 397) = 2.15, *p* < .05		
Constant	2.30(0.20)***	[1.91, 2.70]
Leader's group membership (Group)[a]	-0.17(0.27)	[-0.69, 0.35]
Presence of norm and goal (Presence)[b]	0.46(0.27)	[-0.08, 0.99]
No. of comments (Comments)[c]	0.01(0.28)	[-0.53, 0.56]
Group × Presence	-0.13(0.39)	[-0.89, 0.63]
Group × Comments	0.33(0.39)	[-0.44, 1.10]
Presence × Comments	0.23(0.39)	[-0.55, 1.00]
Group × Presence × Comments	-0.32(0.55)	[-1.40, 0.77]
Covariate: Order[d]	0.18(0.14)	[-0.09, 0.46]
Outcome variable: Perceived appropriateness, R^2 = .03, *MSE* = 1.78, *F*(8, 397) = 1.78, *p* = .08		
Constant	2.67(0.19)***	[2.29, 3.05]
Leader's group membership (Group)[a]	-0.12(0.26)	[-0.63, 0.39]
Presence of norm and goal (Presence)[b]	0.29(0.26)	[-0.23, 0.80]
No. of comments (Comments)[c]	0.20(0.27)	[-0.33, 0.74]
Group × Presence	0.34(0.37)	[-0.40, 1.08]
Group × Comments	-0.43(0.38)	[-1.18, 0.32]
Presence × Comments	-0.64(0.38)	[-1.39, 0.11]
Group × Presence × Comments	0.40(0.54)	[-0.65, 1.45]
Covariate: Order[d]	-0.04(0.14)	[-0.31, 0.23]
Outcome variable: Perceived effectiveness, R^2 = .05, *MSE* = 2.14, *F*(8, 397) = 2.52, *p* < .05		

Constant	2.61(0.21)***	[2.19, 3.03]
Leader's group membership (Group)[a]	-0.16(0.28)	[-0.72, 0.39]
Presence of norm and goal (Presence)[b]	0.33(0.29)	[-0.23, 0.90]
No. of comments (Comments)[c]	0.20(0.30)	[-0.38, 0.79]
Group × Presence	-0.07(0.41)	[-0.88, 0.74]
Group × Comments	0.46(0.42)	[-0.36, 1.28]
Presence × Comments	0.14(0.42)	[-0.68, 0.97]
Group × Presence × Comments	-0.14(0.59)	[-1.29, 1.01]
Covariate: Order[d]	0.16(0.15)	[-0.13, 0.46]

Note. Standard deviations are presented in parentheses. CI = confidence interval.
[a] Leader's group membership was coded as 0 = ingroup/American, 1 = outgroup/Chinese. [b] Presence of leader's norm and goal was coded as 0 = leader's norm and goal were absent, 1 = leader's norm and goal were present. [c] Number of shaming comments was coded as 0 = one shaming comment, 1 = five shaming comments. [d] Order of presenting the questions about the identification with the leader was coded as 0 = before (viewing the shaming punishments), 1 = after (viewing the shaming punishments).
*** indicates $p < .001$.

Leader's Identification

Lastly, a two-way ANOVA was conducted to examine whether there is an interaction between the order of the identification with the leader (either measured before or after) and leader's group membership, in predicting people's identification with the leader (H4). It was found that the interaction between the order and leader's group membership did not predict people's identification with the leader, $F(1, 402) = 2.17$, $p = .36$, $\eta^2_p < .01$. People's identification with the leader was found to be predicted by the order, $F(1, 402) = 75.06$, $p < .001$, $\eta^2_p = .07$, but not leader's group membership, $F(1, 402) = 7.53$, $p = .09$, $\eta^2_p = .01$. It was found that people's

identification with the leader was lower when it was measured after they have seen the shaming posts, than when it was measured before viewing the posts. In sum, the proposed interaction between the order of identification with leader and leader's group membership predicting people's identification with leader (H4), was not supported.

Salience of Identity

Since it was found that people's identification with the leader was affected by whether it was asked before or after participants have seen the shaming posts, indicating a potential disengagement with the leader, we further examined whether such an influence on people's identification with leader would have downstream effects on people's behavioural intention and attitudes of online shaming. In some previous research, researchers have used reflection tasks to manipulate the salience of an identity, such as asking participants what they have in common with other Australian people to make the group identity more salient (e.g., Haslam et al., 1999). Consistently, in the current study, when participants were provided with an opportunity to reflect on their identity with the leader (i.e., to what extent they identify with and feel positive for the leader) before they have seen the shaming comments, their identity with the leader might become more salient, as opposed to an opportunity to reflect later.

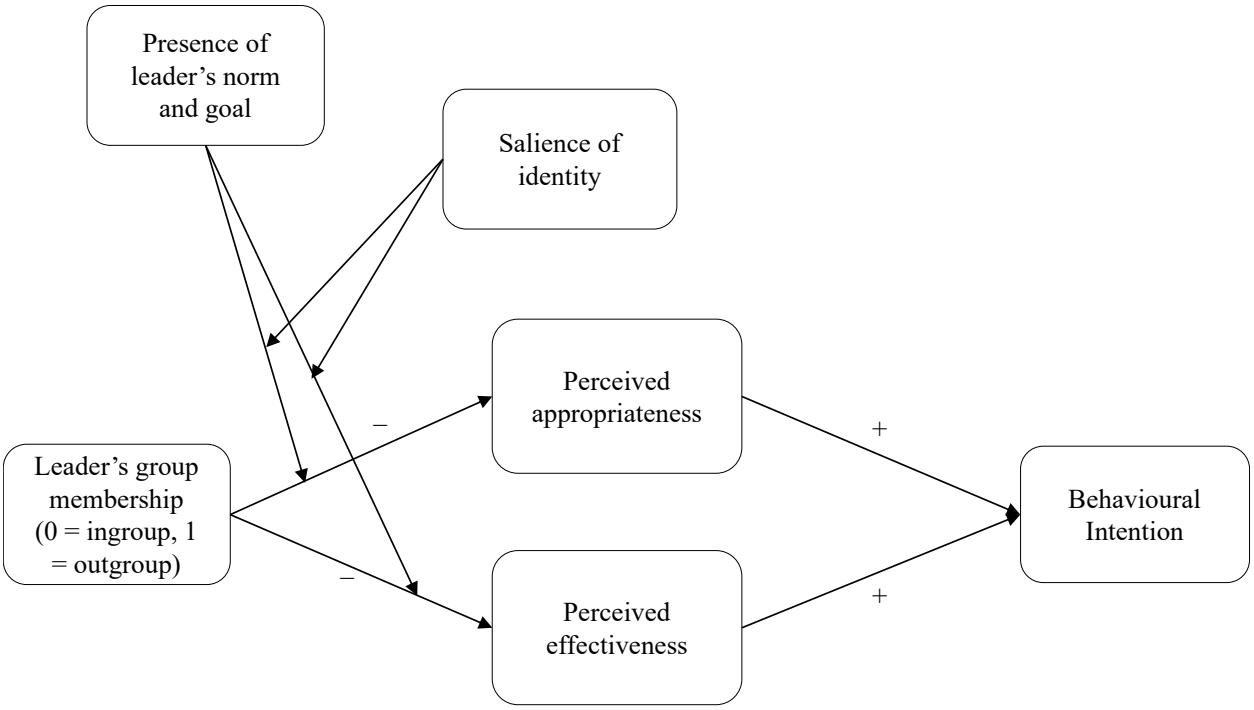
The previously hypothesised relationship between leader's group membership and the presence of leader's norm and goal in predicting people's online shaming attitudes and behaviour, was not supported. However, we suggest this relationship could be intensified by the increased salience of the shared identity with the leader. Based on this rationale, we examined in the post-hoc analysis whether the relationship between leader's group membership and the presence of leader's norm and goal on people's online shaming attitudes and behaviour would be found only when the identity with the leader is more salient, but not when the identity with the

leader is less salient. Moreover, since attitude can predict behaviour (e.g., Wallace et al., 2005), we examined whether people's attitudes about online shaming, specifically their perceived appropriateness and effectiveness of online shaming, would mediate the effect of leader's group membership on people's behavioural intention to support online shaming.

Specifically, as shown in Figure 15, consistent with H1 and H2b, we expected that people who were presented with an ingroup leader would be more likely to perceive online shaming as appropriate and effective than those who were presented with an outgroup leader, and such effects of group membership on attitudes would be stronger when the leader's norm and goal were present than absent. Furthermore, we expected the interacting relationship between leader's group membership and presence of leader's norm and goal on people's attitudes would be stronger when people's identity with the leader is made salient (i.e., when identification with the leader is measured before) than their identity with the leader is less salient (i.e., when their identification is measured after). Additionally, we expected that attitudes (appropriateness and effectiveness) about online shaming could further predict people's behavioural intention to shame, with stronger attitudes being associated with greater intention. In sum, we expected a three-way interaction between leader's group membership, the presence of leader's norm and goal, and the strength of identity salience, on people's attitudes of online shaming (perceived appropriateness and effectiveness; parallel mediators), which in turn, predict people's behavioural intention to shame.

Figure 15

Conceptual Model of the Post-hoc Analysis



Note. Salience of identity was indicated by the order of identification with the leader, with 0 = identity with the leader made more salient (i.e., when identification with the leader is measured before viewing the shaming posts), 1 = identity with the leader made less salient (i.e., when their identification is measured after viewing the shaming posts).

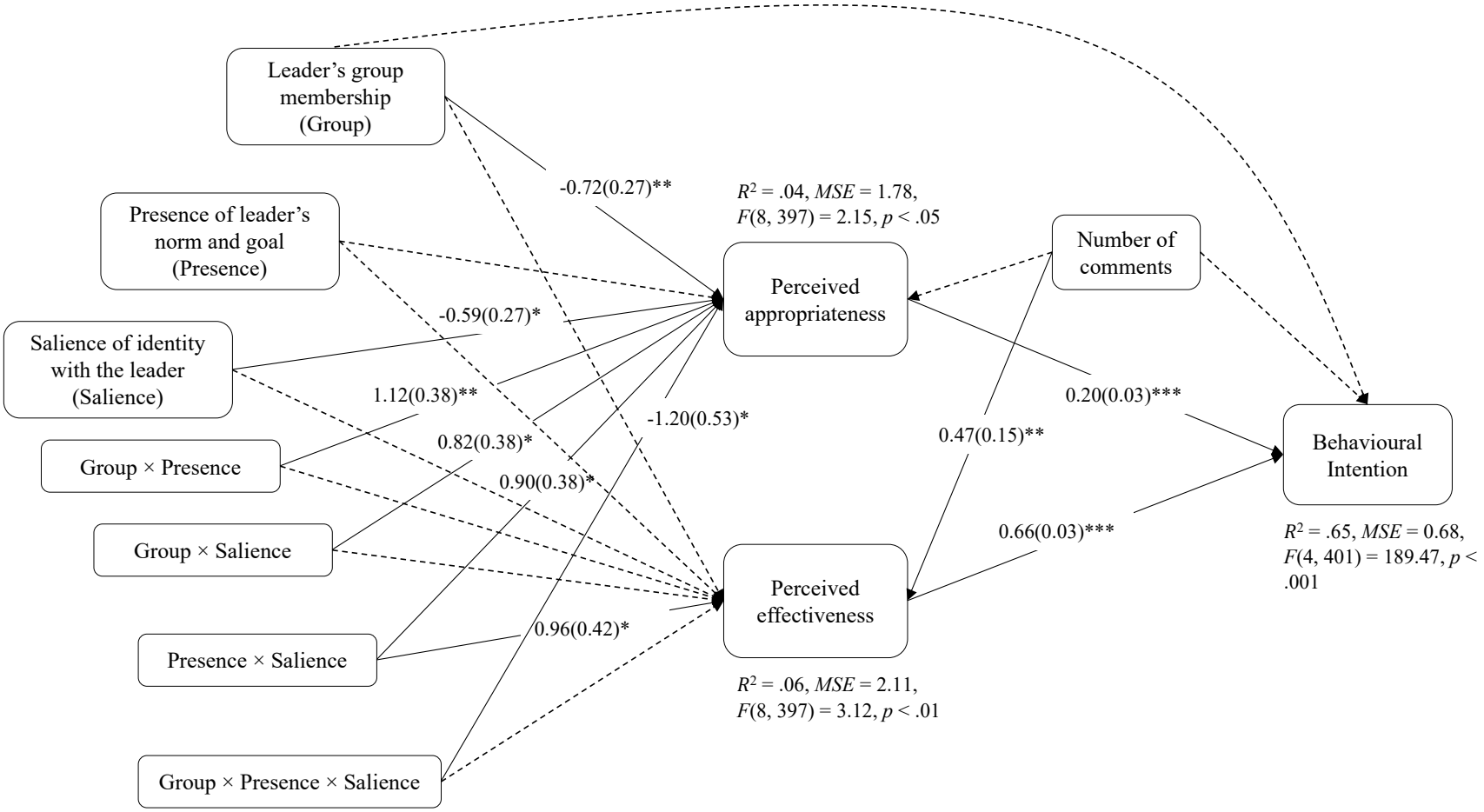
Post-hoc Analysis

A customised PROCESS Model was used to test the moderated moderated parallel mediation model (Hayes, 2017; see Figure 15). Correlation analyses revealed significant positive relationships between the three dependent variables (appropriateness, effectiveness, behavioural intention, all *r*s $> .43$**), irrespective of whether the identity with the leader was made salient. The coefficients of significant predictors are shown in Figure 16, and the remaining coefficients (including the constants) are shown in Table 9.

As shown in Figure 16, when controlling for the number of shaming comments, there was a significant three-way interaction (leader's group membership, the presence of leader's norm and goal, and the salience of identity) predicting perceived appropriateness, $\Delta R^2 = .01$, $F(1, 397) = 5.11$, $p < .05$. Consistent with what we expected, the interaction between leader's group membership and the presence of norm and goal was significant only when the identity with the leader was more salient, $B = 1.12$, $F(1, 397) = 8.63$, $p < .01$, but not when the identity was less salient, $B = -.08$, $F(1, 397) = 0.05$, $p = .83$. However, when identity with the leader was salient, the direction of the interaction between leader's group membership and the presence of norm and goal was contradictory to what we expected. As shown in Figure 17, when the identity was salient and when the leader's norm and goal were *absent*, participants who were presented with the ingroup leader perceived the shaming comments as more appropriate than those who were presented with an outgroup leader (slope = -0.72, $SE = 0.27$, $t = -2.71$, $p = 0.01$). However, such an effect of leader's group membership was not found when the identity with the leader was salient and when the leader's norm and goal were *presented* (slope = 0.40, $SE = 0.27$, $t = 1.47$, $p = 0.14$). Nor was the effect found when the identity with the leader was less salient, irrespective of whether the leader's norm and goal were absent (slope = 0.10, $SE = 0.27$, $t = 0.37$, $p = 0.71$), or present (slope = 0.02, $SE = 0.26$, $t = 0.07$, $p = 0.9$).

Figure 16

Regression Coefficients for the Significant Predictors in the Moderated Moderated Parallel Mediation Model



Note. Coefficients presented are unstandardised regression coefficients for significant predictors. Standard errors are presented in parentheses. See Table 9 for coefficients and standard errors for non-significant predictors. Leader's group membership was coded as 0 = ingroup/American, 1 = outgroup/Chinese. Presence of leader's norm and goal was coded as 0 = leader's norm and goal were absent, 1 = leader's norm and goal were present. Number of shaming comments was coded as 0 = one shaming comment, 1 = five shaming comments. Salience of identity was indicated by the order of identification with the leader, with 0 = identity with the leader made more salient (i.e., when identification with the leader is measured before viewing the shaming posts), 1 = identity with the leader made less salient (i.e., when their identification is measured after viewing the shaming posts). * indicates $p < .05$. ** indicates $p < .01$. *** indicates $p < .001$.

Figure 17

Interactions Between Leader's Group Membership and Presence of Leader's Norm and Goal Predicting Perceived Appropriateness by Salience of Identity with the Leader

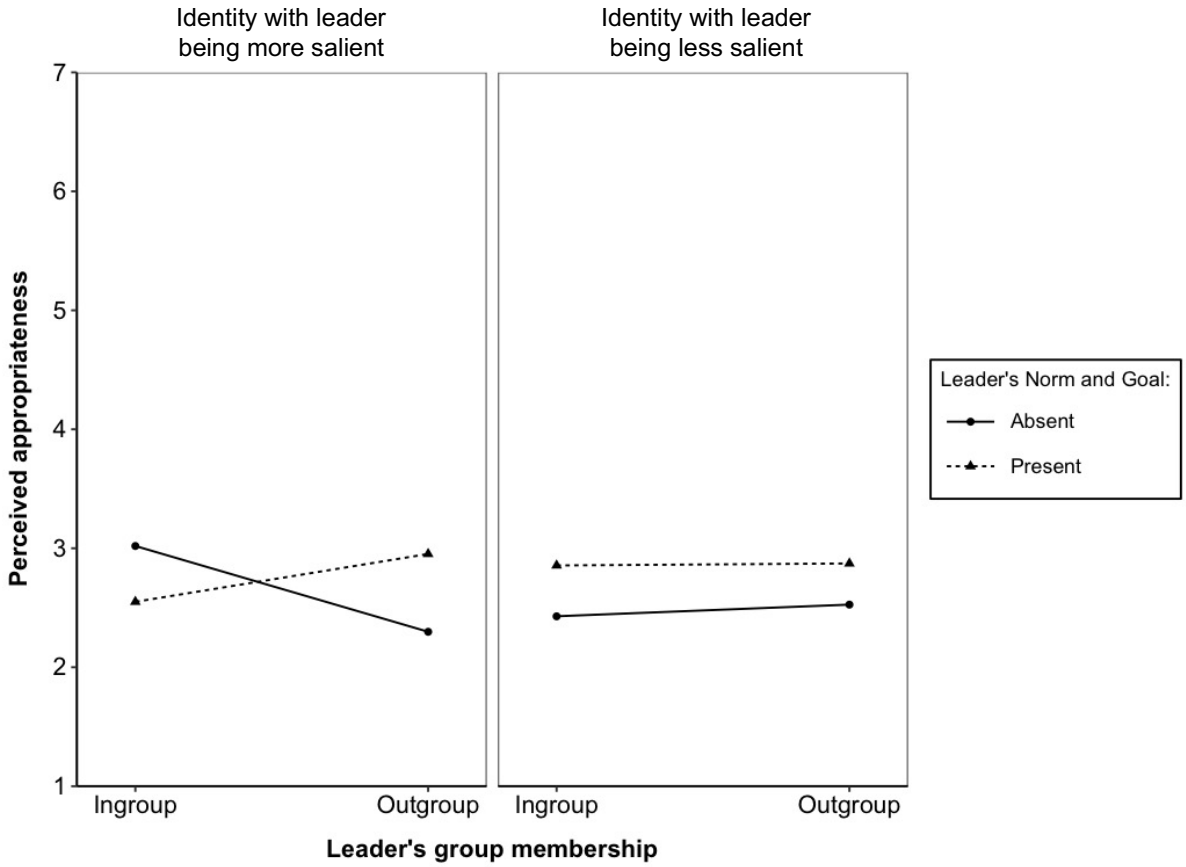


Table 9

Regression Coefficients (and Constants) for Non-significant Predictors in the Moderated Moderated Parallel Mediation Model

Predictor	b (SE_b)	95% CI for b
Outcome variable: Perceived appropriateness		
Constant	3.15(0.20)***	[2.75, 3.54]
Presence of norm and goal (Presence)[b]	-0.47(0.28)	[-1.01, 0.07]
Covariate: No. of shaming comments[d]	-0.26(0.14)	[-0.53, 0.01]
Outcome variable: Perceived effectiveness		
Constant	2.79(0.22)***	[2.36, 3.22]
Leader's group membership (Group)[a]	-0.33(0.29)	[-0.90, 0.24]
Presence of norm and goal (Presence)[b]	-0.10(0.30)	[-0.69, 0.49]
Salience of identity (Salience)[c]	-0.44(0.29)	[-1.02, 0.14]
Group × Presence	0.37(0.42)	[-0.45, 1.19]
Group × Salience	0.76(0.41)	[-0.05, 1.57]
Group × Presence × Salience	-0.96(0.58)	[-2.10, 0.19]
Outcome variable: Behavioural intention		
Constant	0.20(0.12)	[-0.04, 0.43]
Leader's group membership	-0.14(0.08)	[-0.30, 0.03]
Covariate: No. of shaming comments[d]	-0.05(0.08)	[-0.22, 0.11]

Note. Standard deviations are presented in parentheses. CI = confidence interval.
[a] Leader's group membership was coded as 0 = ingroup/American, 1 = outgroup/Chinese. [b] The presence of leader's norm and goal was coded as 0 = leader's norm and goal were absent, 1 = leader's norm and goal were present. [c] Salience of identity with the leader was coded as 0 = identity being more salient (identification with the leader measured before viewing the shaming punishment), and 1 = identity being less salient (identification with the leader measured after viewing the shaming punishment). [d] The number of shaming comments was coded as 0 = one shaming comment, 1 = five shaming comments.
*** indicates $p < .001$.

As shown in Figure 16, the proposed three-way interaction (leader's group membership, the presence of leader's norm and goal, and the salience of identity) was not a significant predictor of perceived effectiveness, $\Delta R^2 = .01$, $F(1, 397) = 2.71$, $p = .10$. However, it was found that the interaction between leader's norm and goal and the salience of identity was significant. We then ran a linear multiple regression to further examine this two-way interaction. It was found that the interaction term between the presence of leader's norm and goal and the salience of identity was no longer a significant predictor of perceived effectiveness (as shown in Table D6 of Appendix D).

Figure 16 further showed that people's perceived appropriateness and effectiveness positively predicted their behavioural intention to support the online shaming campaign. The indirect effects via perceived appropriateness and effectiveness are presented in Table 10. Consistent with the found three-way interaction predicting perceived appropriateness, the indirect effect of leader's group membership on people's behavioural intention via perceived appropriateness was significant only when the identity with the leader was made salient and the leader's norm and goal were absent. This significant, negative indirect effect shown in Table 10 suggests that when the identity with the leader was salient and when the leader provided no information about the norm and goal, people who were introduced to an ingroup leader were more likely to support the campaign via appropriateness than when it was an outgroup leader. The index of moderated moderated-mediation was equal to −.58, bootstrap $SE = 0.29$, bootstrap $CI_{95\%} = [-1.17, -0.01]$, further revealing that the effect of leader's group membership was significant in the moderated moderated-mediation relationship. However, the mediating role of effectiveness was not supported. As shown in Table 10, none of the indirect effects of leader's group membership on people's behavioural intention via effectiveness was significant, nor was the index of moderated moderated-mediation, index = −.63, bootstrap $SE = 0.39$, bootstrap $CI_{95\%} = [-1.41, 0.09]$. In sum, the proposed moderated

moderated parallel mediation (perceived appropriateness and perceive effectiveness) model was only partially supported.

Table 10

Tests of Indirect Effects of the Leader's Group Membership via Parallel Mediators (Perceived Appropriateness and Effectiveness) on Behavioural Intention by Conditions

Conditions	*IE*(*SE*)	95% CI
Via perceived effectiveness by conditions (Presence × Salience)		
Absent, More salient	-0.22(0.17)	[-0.56, 0.11]
Absent, Less salient	0.29(0.18)	[-0.07, 0.64]
Presented, More salient	0.03(0.23)	[-0.42, 0.48]
Presented, Less salient	-0.10(0.18)	[-0.46, 0.25]
Via perceived appropriateness by conditions (Presence × Salience)		
Absent, More salient	-0.14(0.06)	[-0.26, -0.04]
Absent, Less salient	0.02(0.05)	[-0.08, 0.13]
Presented, More salient	0.08(0.05)	[-0.03, 0.19]
Presented, Less salient	0.003(0.05)	[-0.11, 0.11]

Note. *IE* = indirect effect; *SE* = standard error; CI = confidence interval. Presence represents the presence of leader's norm and goal. Salience represents the salience of identity with the leader, with more salient indicating identification with the leader measured before viewing the shaming punishment, and less salient indicating identification with the leader measured after viewing the shaming punishment. Leader's group membership was coded as 0 = ingroup/American, 1 = outgroup/Chinese.

Discussion

Although the manipulation of leader's norm was successful in the current study, the manipulation of leader's goal was not successful. For the manipulation of leader's group membership, it was successful for participants who were asked about their identification with the leader before viewing the shaming posts, but not for those who were asked after viewing the shaming posts. It was also found that none of the proposed hypotheses was supported. We found that none of the main effects or the hypothesised interactions predicted the outcome variables (i.e., appropriateness, effectiveness, and behavioural intention of online shaming). These include the effects of the leader's group membership, the presence of leader's norm and goal, and the number of shaming comments, as well as the interactions between these variables. We also did not find support for the last hypothesis, as people's identification with the leader was not predicted by the interaction between the order of the identification with the leader (either measured before or after viewing the shaming posts) and leader's group membership.

We further proposed that the order of identification with the leader, either the participants were asked to reflect on their identity was the leader before or after viewing the shaming posts, could indicate different levels of identity salience. For people who were provided with an opportunity to reflect on their identity with the leader before viewing the shaming comments, their identity with the leader might become more salient, as opposed to those who reflected later. Accordingly, we tested additional hypotheses in the post-hoc analysis. Partly consistent with the hypotheses, we found that when the identity was salient (i.e., when identification was measured before) and when the leader's norm and goal were absent, the leader's group membership was found to predict people's perceived appropriateness (but not effectiveness) of online shaming, with downstream effect on their willingness to support shaming behaviour. Specifically, people who were introduced to an

ingroup leader reported stronger perceived appropriateness and greater behavioural intention to shame, than those who were introduced to an outgroup leader. Such effects of group membership on people's behavioural intention via perceived appropriateness were not found when the identity with leader was less salient (i.e., identification being measured later), or when the leader's norm and goal were present.

However, inconsistent with our expectation that the effects of group membership on attitudes would be stronger when the leader's norm and goal were present than absent, we found that the effect of leader's group membership on perceived appropriateness only when the leader's norm and goal was absent, but not when it was present. One explanation could be that people who were presented with the leader's norm and goal might have experienced reactance to the leader's norm and goal. Psychological Reactance Theory (Brehm, 1966) claims that persuasive messages can produce negative arousal in people when they feel that their freedom to respond in their own preferred way is threatened. As a consequence, people's commitment to engage with the message (e.g., following the leader's norm and goal in this study) could be undermined. Consistent with this explanation, we found that participants perceived the leader's goal as less noble when the leader's norm and goal was present, compared to the absence of such information. Indeed, people generally showed a reluctance to support online shaming as well as perceived online shaming as inappropriate and ineffective. As we will explain in more detail below, along with what we found in Study 3, these findings suggested that people tend not to be blind followers of the leader when it comes to online shaming.

General Discussion

The studies presented in the current chapter aimed to examine whether people's engagement in online shaming punishment could be guided by their identification with the leader (i.e., leader's group membership) as well as the leader's influences (i.e.,

presence/absence of leader's norm and goal). Results across two studies provided consistent evidence that people generally do not prefer using online shaming for behavioural deterrence, irrespective of the leader's group membership: In Study 3, participants were asked to engage in online shaming directly and the shaming options were at a high level of hostility for the last two trials, which might have caused people to feel reluctant to engage in online shaming. Despite this issue was further addressed in Study 4 with a more indirect behavioural measure that did not ask people to choose or contribute a shaming comment, people still showed a reluctance to engage in shaming, and generally perceived online shaming as ineffective and inappropriate.

Importantly, it was found that people who were presented with the leader's norm and goal perceived the leader's goal as less noble, compared to those were not presented with such information. This finding could be explained by a felt pressure to engage in shaming due to the presence of leader's norm and goal, particularly because people's views about online shaming differed from the leader's. Indeed, when participants were asked about their view on the public health campaign, some indicated that although they agreed with the goal of the campaign, they were not comfortable with the idea of using online shaming as a punishment (e.g., "This campaign is just targeting people and not fairly"), and/or questioned its effectiveness (e.g., "I don't think 'shaming' people into compliance is effective vs. other methods"). Such incompatible views (and the underlying values) could have posed a threat to people's feelings of oneness with the leader (i.e., an identity threat; Ellemers et al., 2002), which could be further examined in future research.

Together, the findings of people showing reluctance to shame suggest contradiction to the destructive obedience to authority proposed by Milgram (1963, 1974). Rather, consistent with the study results that did not receive enough attention in Milgram's interpretation, there was always some non-compliance existed, and the non-compliance increased along with the

increasing severity of the punishment (Russell, 2011). Therefore, aligning with Reicher et al.'s (2012) reconceptualisation, it may be that people kept a sense of agency by holding a critical view and showing behavioural resistance towards online shaming. Indeed, our findings suggested that the participants generally do not prefer using shaming comments to punish others or deter others' behaviour when it comes to addressing online hostility or violation of COVID-19 public health guidelines. Therefore, leaders who demonstrate undesirable behaviour such as online shaming in these contexts might not unsuccessful in mobilising others.

People's online shaming engagement, however, can still be influenced by an identity leader. It was found in the additional analysis that when the identity was salient and when the leader's norm and goal were not presented, people who were introduced to an ingroup leader were less likely to find online shaming inappropriate and were more likely to engage in online shaming, compared to when introduced to an outgroup leader. This suggests that despite people might have their own understandings on the practice of online shaming (i.e., being negative in general, as reflected from the shown resistance), when considering a specific case of online shaming, people can perceive it as less inappropriate in the situation and be more willing to engage in shaming. Especially, the study shows that under certain circumstances, social identification plays a crucial role in this process and shapes people's online shaming attitude and engagement.

These results provide both theoretical and empirical support for understanding online shaming as a group behaviour. Through conceptualising online shaming as a group-based punishment, this study extended on the engaged followership model (e.g., Haslam et al., 2015) and further applied it to the context of online shaming. Previously, online shaming was only understood as an aggregated punishment (e.g., Sawaoka & Monin, 2018, 2020), with the idea of it being a group-based punishment remains underexplored. To my knowledge, this

study serves as the initial examination of the role of identity leader in mobilising online shaming as a group-based punishment. Furthermore, consistent with the suggestion on exploring online moral expressions through a lens of social identity (Brady et al., 2020; Marwick, 2021), the current study provided one of the first experimental examination of the group-based motives that drive online shaming.

Specifically, it provides a possible explanation for why some people still engage in online shaming, despite that the public generally shows a negative view towards online shaming. Consistent with the previous research (Oakes, 1987), identity salience can be a prerequisite for a social identity to successfully influence people's attitude and behaviour. Therefore, when the audience of shaming belongs to an ingroup and when the shaming happens in a relevant context for the group identity to be made salient, online shaming might be perceived as less inappropriate than when these conditions about identity are not met. Through evidence for that online shaming is shaped by social identity as a group behaviour, we help point to directions for future research on the reasons behind the ongoing debates surrounding this practice (Muir et al., 2021; Vogels et al., 2021; see also Chapter 3). We suggest that future research can continue to examine whether, and to what extent, people's different understandings on a specific case of online shaming (e.g., the perceived appropriateness) are informed by their group membership. For example, future research can compare between how ingroup and outgroup members might respond differently to the same norm violation.

Our findings also suggest that online shaming can have unintended – or unforeseen – consequences for those who resort to this practice for mobilisation. Although the group leader could influence people's attitude and behavioural intention about online shaming, the leader's mobilisation was not always effective. This is consistent with research indicating that there may be competition among leaders within a group (Blackwood & Louis, 2017). In other

words, the social identity leader that people "prefer" is context-dependent, with certain situations in which people favour leaders who negotiate with the outgroup. Therefore, when online shaming is imposed by the leader as an explicit norm, it could pose threats (e.g., potential backlashes) to those who are asked to engage. Previous research found that explicit norm-based interventions can pose threats to one's autonomy, values, as well as morality (e.g., Bergquist & Nilsson, 2016; Bosson et al., 2020; for review see Wenzel & Woodyatt, 2025). Especially, when one defines their identity as the opposite to the imposed norm enforcement (e.g., the norm to shame) and/or when they do not see the leader (e.g., who initiates online shaming) as representing the group, the norm enforcement could lead to the bolstering of one's oppositional identity based on resistance to the imposed norm (Turner, 2005). Therefore, the findings of this study suggest that online shaming as an explicit strategy for achieving a collective goal should be used with caution, due to the accompanied risks of not achieving the intended outcomes and potentially creating resistance.

Limitations and Future Directions

One limitation of the current research is that the shaming comments shown to the participants ranged from moderately hostile to extremely hostile, which might not have provided an accurate representation of what happens in a real-life shaming event. Past instances of online shaming have featured various types of comments, ranging from sarcasm and jokes to passing judgment and releasing personal or private information, as well as more benign, educational remarks (Basak et al., 2019; see also Chapter 3). Future research could explore whether people perceive online shaming as more or less appropriate when exposed to these different types of comments, and whether this would affect their online shaming engagement. Specifically, it can be explored along with the different goals that motivate people to engage in shaming. For example, whether people who shame to raise others' awareness (i.e., the goal of creating change) differ from those who shame to punish others

and seek justice, in terms of the types of shaming comment that they contribute and/or support.

Another limitation of the current research is that the manipulation of social identification with the leader used in the current research, which was based on the leader's nationality. Although the manipulation was successful in Study 4 among participants who reported their identification with the leader before viewing the shaming posts, it was unsuccessful among those who did so after viewing the posts. One possible explanation for this between-group discrepancy is that participants who first saw the shaming posts were less willing to identify with the leader, reflecting a general reluctance to associate with or endorse online shaming. However, an alternative explanation is that the nationality-based identity manipulation may not have been sufficiently relevant to the context of online shaming. For example, it was found in the US that Democrats and Republicans had different views on the COVID-19 pandemic and differed in their adherence to the COVID-19 guidelines (Pew Research Center, 2020; Gollwitzer et al., 2020). Hence a manipulation of leader's group membership based on political group memberships could be more suitable in the context of online shaming relating to COVID-19. Though it is important to note that the Study 4 was conducted very early in the timeline of COVID, when highly polarised responses were only in their infancy as the world was still coming to, with a general awareness of the virus as emerging from China.

Finally, the current study found that in an experimental setting, identity leaders can influence people's online shaming of those who violate COVID-19 public health guidelines under certain circumstances (i.e., when the identity with the leader being salient and the leader's norm and goal being less explicit). However, we suggest that these findings need to be interpreted with caution, given that they were not hypothesised prior to the conduct of the study. Future research is needed to further examine whether and how identity leaders

influence people's online shaming engagement, especially in other contexts involving different types of norm violations.

Furthermore, it remains unexplored how identity leaders might emerge during online shaming. In addition to cases where one's identification with the leader converges with an existing group identity (as shown in the current research), opinion-based groups can form through online discussions (McGarty et al., 2009; Thomas & McGarty, 2009), and a role of leader can emerge from this process. Indeed, analysis on real-life online shaming found that shaming involved the formation of distinct communities as well as the involvement of both intragroup and intergroup interactions (Chapter 3). We suggest future research could examine explicitly how opinion-based leaders emerge during an online shaming event, as well as how the responses of a leader can shape the people's subsequent shaming engagement. Specifically, building on the existing research (Chapter 3), we suggest that using archival data to uncover the patterns and structures among user's interactions within a community (intragroup interaction) as well as interactions between different communities (intergroup interaction) (see also Cabrera et al., 2021; Maher et al., 2020), could be a fruitful way to further examine the emergence of leader and their influences on online shaming.

Conclusion

The current chapter presents of two studies that examine how an identity leader can mobilise others to engage in online shaming. Across both studies, participants generally held negative views toward the practice of online shaming. Additional analyses of Study 4 further revealed that people's attitude and engagement in online shaming can be influenced by an identity leader under certain conditions (i.e., when the identity is salient and when the leader's punitive norm and noble goal are not explicitly imposed). These studies offer one of the first empirical investigations of online shaming as a group behaviour that can be mobilised through shared social identities. However, future research is needed to further

examine online shaming as a group behaviour and to explore the role of the identity leader in shaping group processes. Overall, the findings highlight that online shaming may have unintended or unforeseen consequences for those who resort to this practice for mobilisation.

CHAPTER 5. General Discussion

Online shaming as a phenomenon has attracted widespread media and public attention. This thesis has contributed a theoretical framework and empirical evidence for *why* people engage in online shaming, exploring online shaming as a group-based phenomenon that can be utilised in the pursuit of multiple group goals. This research not only contributes to the research in online shaming but also to a much broader literature regarding justice, online activism, and online group behaviours. Across 4 studies, I have provided evidence that people who engage in online shaming can be driven by more than just the individual's intent to harm – or, literally "bring shame upon" – a wrongdoer. I found evidence that online shaming is characterised by group dynamics and motivated by multiple discrete goals (i.e., punishing the perceived wrongdoer, deterring the perceived wrongdoing, seeking social acknowledgement, and creating change). I also observe that when a common goal is shared among a group, online shaming may serve to fulfill one's identity-based needs, which potentially benefits the ingroup and/or the society at large. In the present chapter, I provide a summary of the key findings for each of the studies along with the theoretical contributions. I then present a discussion of the practical implications. At last, I consider the strengths and the limitations of the current research and suggest directions for future research to further our understanding of online shaming, with a focus on the gaps and novel findings that emerged across the studies. These findings not only provide researchers the insights on how online shaming can be conceptualised and approached in future studies, but also inform different social actors (e.g., policy makers and social media platforms) on how to respond to future shaming occurrences.

Online Shaming Driven by Group Processes and Shared Goals

A key contribution of my thesis was conceptualising online shaming as more than just an exchange that occurs between individuals. I argued that online shaming should be

understood as a group behaviour, which is shaped by group processes and shared goals. This perspective shifts the focus from individual-level explanations to broader, group-level motives that can drive people's engagement in shaming. In Chapter 2, I examined the multidisciplinary literature on online shaming in a scoping review, which involved a thematic synthesis of the existing literature on why people engage in online shaming (Study 1). I reviewed 94 articles of various disciplines, such as psychology, sociology, media and communications, law, and computer science, to provide a holistic view of the multidisciplinary literature.

The scoping review revealed that online shaming can be driven by multiple goals that people have. Such goals capture the various motives that people may have when they participate in online shaming. The goal of punishing the perceived wrongdoer (the punishment goal) reflects the motivation of actively seeking retribution against the perceived norm violator. The goal of deterring the perceived wrongdoing (the deterrence goal) focuses on the prevention of future occurrences of the norm violation that would potentially be conducted by either the perceived wrongdoer or others. The goal of seeking social acknowledgement describes one's desire to seek recognition and approval from others (particularly, other ingroup members). And lastly, the goal of creating change identifies online shaming as a strategy to mobilise others to one's cause, which is also often shared collectively within a group.

Given that the literature on online shaming is multidisciplinary and employs diverse methodologies, the scoping review offered an opportunity to synthesise the heterogeneous literature. Indeed, I found in the scoping review that recent literature on online shaming has become more sophisticated, with a greater acknowledgement that online shaming is a behaviour that involves not only the intent to harm an individual, but also other goals – including deterring the perceived wrongdoing, seeking social acknowledgement, and creating

change. The synthesis of these heterogeneous literature also generated insights for future studies to explore online shaming in a greater depth. I suggested that further research is necessary to test the identified goals and the group-based nature of online shaming – for example, future research could address whether online shaming can be a form of group-based interactions that shape group identities and mobilise others to create change.

The psychological goals identified through the initial scoping review (Chapter 2) guided the subsequent empirical work in my thesis (Chapters 3-4). These findings not only provided a basis for the conceptualisation of online shaming, but also informed the design of the subsequent empirical studies and the interpretation of their data. Chapter 3 provides a direct empirical test of the relevance of the goals identified by the scoping review (Study 4). In Chapter 3, I further investigated why people engage in online shaming by verifying whether the identified goals existed in a real-life online shaming event on Twitter. Through collecting and analysing real-life comments from Twitter (N = 5005), I found evidence that online shaming can be driven by both collective goals and group processes. The online shaming event spanned 4 days on Twitter, which provided me an opportunity to examine how the identified goals progressed over time.

This pandemic shaming event, being the focus of Chapter 3, began when a health minister publicly criticised a doctor who was seen by some as violating COVID-19 public health guidelines, while others believed the doctor's actions were appropriate. I found that two discrete groups had formed based on their understandings on who the perceived wrongdoer was: One group formed via shaming the doctor, following the health minister's call-out of the doctor. Another group formed via shaming the health minister and supporting the doctor (the group appeared to be mainly of healthcare workers). The results from Topic Modelling showed evidence for the goals identified from the scoping review (Chapter 2). I found that each of the four goals (punishing the perceived wrongdoer, deterring the perceived

wrongdoing, seeking social acknowledgement, and creating change) was present in both groups, whether shaming the doctor or shaming the health minister. Moreover, these goals became more prominent as group-based expressions intensified over time, reflected in both increased support for the ingroup and stronger derogation of the outgroup.

On the other hand, the two groups ("shaming the doctor" and "shaming the health minister") exhibited distinct characteristics. Besides their different understandings on who the wrongdoer was, the groups also evolved into different norms and practices as online shaming progressed. Specifically, the group who shamed the health minister demonstrated a system-challenging norm (Jost et al., 2017; Osborne et al., 2019), by collectively requesting the health minister to address the systemic issues faced by healthcare workers. In responding to the outgroup's behaviour of shaming the doctor, the healthcare workers manifested a goal of creating change, by showing further support for their ingroup and engaging in collective action (i.e., sharing petitions). Conversely, the group who shamed the doctor responded negatively to the doctor-supporting group and defended the health minister's act of calling out the doctor. In doing so, they justified the existing system (i.e., the healthcare system) and contested the validity of the other group's collective action (Jost et al., 2017; Osborne et al., 2019). This supports my argument that goals and group processes can interact, as for each of the groups, the collective goals that they pursued were shaped by their interaction with the other group.

The Twitter analysis (Chapter 3) provided a test of the propositions that online shaming can be driven by both psychological goals and group processes. The test was rich in ecological validity and provided "real world" evidence for the goals identified in the scoping review. It also illustrated the group processes and intergroup interactions that underpin online shaming. Furthermore, I suggested that online shaming could be understood as a group discussion where shared norms and identities are allowed to form (Smith et al., 2015). The

two groups in the Twitter analysis used online shaming for both supporting and opposing social change in ways analogous to online activism.

Taken together, the first two studies made a crucial theoretical contribution by supporting the claims that online shaming can be studied as a group behaviour driven by both group processes and shared goals, individually and in combination. Previous research acknowledged the involvement of a collective of individuals in engaging in online shaming (Brady et al., 2021; Gruber et al., 2020; Johnen et al., 2018; Sawaoka & Monin, 2018, 2020), and that in certain cases, online shaming could entail collective goals (Blitvich, 2021, 2022; Gao, 2013). However, online shaming had not been examined explicitly as a group behaviour, neither were the goals of online shaming understood as being shaped by one's group identity or the intergroup interactions. I addressed this gap by bridging between the diverse understandings on why people engage in online shaming. By applying the integrated understanding to a naturally occurring shaming event on social media, I found additional evidence supporting the group-based nature of online shaming. Therefore, I underscore the value of using social identity approach (Tajfel & Turner, 1979; Turner, 1985) to understand online shaming. Especially, online shaming can both be shaped by one's group identity and serve as a dynamic intergroup process itself through which group identities are formed and (re)defined.

A novel contribution of this research lies in demonstrating that online shaming should be understood as more than a "pile-on" that is impulsive and disordered. Instead, online shaming can be shaped by group identities and intergroup interactions that are constantly evolving. This argument aligns with previous research examining the ways in which crowd behaviour can spontaneously emerge from salient social identities and intergroup interactions, with the latter creating a context for social identities to alter and transform (Drury & Reicher, 1999; Drury & Reicher, 2005; Reicher, 1984). In other words, online shaming, like crowd

behaviour, is not atavistic and disordered – rather it is governed by group processes with individuals pursuing relevant (sometimes diverging) group goals. The current thesis extends this understanding to the behaviour of online shaming that occurs on social media (see also Reicher et al., 1995; Spears et al., 2002).

Online Shaming Mobilised by the Role of Leader

Another key contribution of my thesis was to extend the focus on group processes to examine the role of leader in influencing people's online shaming engagement. The third and the fourth studies presented in Chapter 4 experimentally examined online shaming through the lens of a leadership-followership dynamic. While online shaming was suggested to be shaped by group-based motives and driven by goals (see Chapters 2-3; also, Blivich, 2022; Brady et al., 2020), empirical studies that explicitly examined this dynamic remain limited. The experiments presented in Chapter 4 address this gap by empirically examining online shaming as a group-based punishment that can be used by an identity leader to mobilise others via the presence of a group norm and a collective goal. To the best of my knowledge, these are also the first studies that empirically test the role of leadership in shaping online shaming behaviour.

Specifically, I investigated in Study 3 (as presented in Chapter 4) whether people would engage in online shaming to punish others when they are mobilised by an identity leader (i.e., someone who belongs to the ingroup and presents a punitive norm and a noble goal to justify the shaming). I reasoned that the identity leader would be particularly influential in engaging people to shame others due to the influences of the shared social identity on one's behaviour, when compared to a non-mobilising leader who belongs to an outgroup and does not provide information of the punitive norm or the noble goal. Following Study 3, I further disentangled the effects of mobilising identity leadership by examining whether and how people's identification with the leader interacts with the leader's influence

(via the presence of norm and goal) in predicting people's intention to shame (Study 4). For both studies, I used experimental methods to manipulate the leader's identity by nationality.

The studies consistently showed that people generally did not prefer using online shaming to punish others and held a negative view towards online shaming. In Study 3, the participants became more disengaged as the online shaming punishments became more hostile, even when they were introduced to a mobilising leader. In Study 4, the participants again showed reluctance to engage in online shaming, and perceived online shaming as being inappropriate and ineffective. Furthermore, the manipulation of the leader's noble goal was unsuccessful. Additionally, in Study 4, when their identity with the leader was salient, the participants reported lower identification with the leader when the leader's norm and goal were present than absent, which showed a reactance to the leader's punitive norm and noble goal.

I reasoned that these findings align with Frimer and Skitka's (2018) research on Montagu Principle in terms of the costs of incivility. Montagu principle refers to the idea that being civil can lead to benefits to one's reputation. Frimer and Skitka found that being civil either helped or did not affect political leaders' reputation. Incivility, however, costed political leaders' reputation, as the leaders were perceived as less warm and less favourable when they made uncivil remarks than civil remarks. The findings of my experimental studies are consistent with these results by showing that using online shaming as punishment (i.e., being uncivil) can provoke backlash, instead of offering benefits, to one's leadership. Indeed, it was suggested that expression of moral emotions (especially moral outrage) can be dominated by those who have more extreme views (Van Bavel et al., 2024). However, when people were asked about what they would like to see goes viral online, they preferred content that was less divisive, less hateful, and involved less intense emotions (Rathje et al., 2024). The cost of

incivility, along with what people prefer to see online, provides an explanation to why people generally did not prefer using online shaming to punish others, as shown in Study 3 and 4.

Nevertheless, the additional analysis in Study 4 further showed that the social identity still played a role in shaping people's attitude about online shaming. When the identity with the leader was salient and when the leader's norm and goal were absent, the ingroup leader was more influential than the outgroup leader in leading people to view online shaming as less inappropriate, which subsequently led to a greater intention to engage in shaming. These findings support the claims of the engaged followership model where people are not blind followers but engaged followers of the leader, especially when the behaviour involves hurting others (Haslam & Reicher, 2017; Haslam et al., 2015). Belonging to the same social category indeed provided the leader a basis to influence to some extent people's perceived appropriateness of online shaming, and subsequently, their intention to shame. It is worth noting that, people still viewed online shaming negatively even when they were introduced to a leader who belongs to the ingroup. Participants showed reactance to the leader's punitive norm and noble goal as well as reluctance to engage in shaming, instead of compliance with the leader. These findings are consistent with the previous research that suggested explicit norm enforcement accompany risks and pitfalls (see Wenzel & Woodyatt, 2025, for a review), such as when it is used by a leader who might not fit in the social context (Blackwood & Louis, 2017; Haslam et al., 2023).

The experimental studies presented in Chapter 4 makes a significant theoretical contribution by applying the established social identity theories (i.e., theories on punishment and followership-leadership dynamic) to address emerging questions in the field of online shaming. This investigation specifically addresses the group processes and one of the psychological goals that might drive online shaming engagement (i.e., the goal of punishing the perceived wrongdoer). Future research may adopt the social identity approach and

explore other areas of literature to investigate the other goals identified in my work on this topic (e.g., Chapter 2). For example, theories on collective action and allyship (Barron et al., 2023; Bliuc et al., 2007; Thomas et al., 2012; Traversa et al., 2023) can be applied to examine people's online shaming behaviour that is driven by the goal of creating change.

Taken together, the findings from Chapters 2 and 3 underscore the importance of understanding online shaming as a group-based behaviour. Despite some research of why individuals engage in online shaming (e.g., Ge, 2020; Hou et al., 2017; Muir et al., 2023), little attention has been paid to the group-level motivations behind it. This thesis addresses the gap by offering a theoretically informed, integrated perspective that frames online shaming as being driven by shared goals and group processes. Findings from Studies 1–4 support the view that online shaming is indeed shaped by and reflects group dynamics, including both intragroup processes and intergroup interactions.

Practical Implications of the Studies

The current thesis offers practical insights for addressing the complexities of online shaming in real-world settings. There has been a growing interest in understanding and mitigating the negative consequences of online shaming, such as the recent trends in classifying, detecting, and reducing online shaming based on its "toxicity" (Bodaghi et al., 2023; Li et al., 2024; Record & Miller, 2022). A similar concern was also articulated at a policy level. For example, the Singaporean government considered banning cancel culture (which often involves online shaming), due to their negative impacts such as online harms (Jalah, 2023). However, my studies showed that, while online shaming entails the goal of punishing individuals, it can also advance a group's agenda in ways that reflect the desire for justice and/or social change. Therefore, it is important to understand that there are multiple discrete motives underlying online shaming. In the following paragraphs, I provide suggestions on how my results can inform different social actors on how to address online

shaming. These social actors include the press, social media platforms, as well as policymakers.

First of all, I would like to emphasise the role of the press (i.e., traditional media – newspapers, television) in the process of online shaming. The Twitter analysis showed that the shaming of the doctor started from news sharing. As online shaming progressed, the press also played a role in further exposing and sharing personal information of the doctor (e.g., the doctor being the father of a celebrity). Consistent with previous research (e.g., Trottier, 2018), these findings showed that traditional media can shape or even mobilise the public condemnation of individuals to some extent, such as via exposing the details of the perceived wrongdoing and/or the wrongdoer's personal details. Therefore, it is important for the press to consider their own contribution in the process of shaming. I suggest that newspapers, televisions and other traditional media outlets should give more consideration to the trade-offs between infringing individual rights (such as privacy) and serving the public interest. For instance, during the COVID-19 pandemic, sharing a doctor's travel history and workplace location might aid in controlling the spread of the virus. But disclosing irrelevant personal information (such as details about his daughter) could lead to his identification, which should be cautioned or even avoided.

Secondly, social media platforms play a particular role in directly influencing people's engagement in online shaming. Certain affordances of social media platforms can facilitate and increase the polarisation of groups with opposing views (Bliuc et al., 2021). For example, social media platforms often use algorithms designed to incentivise user interaction and sharing, such as recommending contents likely to induce anger (Brady et al., 2021). Indeed, individuals often encounter norm (and value) violations on social media (Brady et al., 2020), which typically trigger people to shame others in the first place (Chapter 2). To address online shaming and its related issues, some social media platforms use strategies that can

detect and reduce “toxicity” in expressions (Anjum & Katarya, 2023; Basak et al., 2019). However, as has been demonstrated in my thesis, not all types of online shaming can be captured with “toxicity”. Rather, online shaming can involve motives such as reinforcing moral norms and creating change (Chapter 3), and take various forms such as disparaging jokes or remarks.

I acknowledge the importance of tackling the negative consequences of online hostility through detection and mitigation. However, I argue that this strategy alone is insufficient to effectively address online shaming. Instead, platforms can supplement the toxicity-reduction strategy with other ways to facilitate civil conversations. For example, research examining Twitter comments about COVID-19 found that the size of users’ social networks and the amount of positive feedback they received were negatively associated with their use of uncivil expressions (Kim, 2020), where users who had more followers and received more positive feedback expressed less incivility. These findings suggest that social media platforms may facilitate civil conversations by encouraging users to diversify their social network and engage positively with other users.

Because online shaming can involve different motives, I suggest that policymakers need to recognise and address online shaming as a distinct phenomenon. This can mean regulations and interventions that not only detect and remove toxicity, but also promote and support civil norms in the online communities. Indeed, norms were found to shape the characteristics of the language used online, such as the use of aggressive wording (Postmes et al., 2000; Rösner & Krämer, 2016). My Twitter analysis also revealed that distinct group norms emerge from different groups that engaged in shaming via both intragroup and intergroup interactions (Chapter 3). Therefore, it is important for the policymakers to understand the group-based nature of online shaming and address it accordingly. For example, future intervention may focus on addressing group norms within online

communities and incorporating intergroup contact to reduce certain types of online shaming (Wachs et al., 2024), particularly those directed at specific social groups such as marginalised communities (Huffman, 2016; Marwick, 2021).

In summary, my findings imply that different social actors (the press, social media platforms, and policymakers) should be cautious when it comes to intervention in addressing online shaming. Social actors need to approach online shaming as a group behaviour with nuances (i.e., in some cases online shaming can be analogous to online activism), while acknowledging their own contribution in the process of online shaming. Especially, future research is still needed to develop specific guidelines on how the press, social media platforms, as well as policymakers can effectively address the issue of online shaming. For example, one interesting question is how social media platforms, each with distinct affordances, can design and implement practices that promote more civil exchanges between users, as a means of supplementing the existing toxicity-reduction strategies.

Strength, Limitations, and Future Research Directions

One key strength of the present thesis is the use of multiple methodologies to collect and analyse data from various sources. I took an open-ended stance by commencing the research with an inductive approach, and then combined the insights from both academic knowledge on online shaming and naturalistic data produced in people's everyday interactions on social media. I used thematic analysis to synthesise the literature included in the scoping review, topic modelling that uses natural language processing techniques to examine real-life shaming comments on Twitter, as well as experiments that capture people's attitudes, behavioural intentions, and actual engagement in online shaming. My approach ensures the practical relevance of the research by triangulating the existing academic knowledge, the naturally occurring data, and the experimental approaches to provide converging evidence on my key claims (Heale & Forbes, 2013).

In particular, I used both "bottom-up" and "top-down" approaches to investigate the goals and group processes that drove people's online shaming engagement. Previously, social psychologists have argued for a need to explain human social actions in consideration of the context in which they occur (e.g., Reicher, 2004; Drury & Reicher, 1999). In Study 2, I used topic modelling, a novel approach that employed text-mining and natural language processing techniques to examine the online comments sourced from a real-life shaming event on social media. Topic modelling makes use of computational methods to uncover the underlying themes or topics within the textual data based on the co-occurrence of words and phrases (Blei & Lafferty, 2007; Finch et al., 2018). This method adopts a "bottom-up" approach and allows for exploring factors and characteristics that could otherwise not be captured in experimental studies, such as the intergroup dynamics that I found to have played a crucial role in influencing people's online shaming engagement.

The use of archival data or naturally occurring data that have not been produced specifically for research purposes, increases the ecological validity of the study findings. As online shaming behaviour may be deemed by some as socially undesirable, participants might be reluctant to show their intention to engage in shaming in experimental settings. However, such an issue can be circumvented by collecting and analysing comments from social media.

The "bottom-up" approach was paired with a "top-down" approach that involved two experimental studies designed to isolate the influences of social identity on online shaming (Study 4) as well as people's actual engagement (Study 3). The findings that people generally do not prefer using online shaming to punish others confirm that this behaviour is perceived as socially undesirable and highlight the methodological challenges of studying it through experimental designs. Nonetheless, this approach was important in empirically testing some of the findings from the topic modelling analysis with strong internal validity.

There are several limitations needed to be acknowledged. Firstly, I did not investigate how certain affordances of specific platforms on social media, as well as the differences among platforms, might affect people's online shaming engagement. People's expression on social media can be influenced by certain situational factors relevant to the design of platforms, such as anonymity and the types of social feedback users receive (e.g., the display of downvotes) (Puryear, 2020; Rost et al., 2016). Such factors might influence the salience of social identity and the interactions between users on the platform (Brady et al., 2021), which might affect the occurrence and/or progression of online shaming. Since online shaming events often happen simultaneously across different platforms, future studies can make a comparison of the same event across platforms and/or communities to examine the potential differences. Future research can also empirically test how the situational factors, such as anonymity, might interact with the group processes in shaping online shaming engagement.

Another limitation of the current thesis points to the use of only experimental studies to examine the role of leader. However, in real-life online shaming events, the role of leader can emerge as online shaming progresses. Future research may use a network analysis approach to address this limitation. It can identify individuals who are more central and influential in shaping people's opinions and behaviours within a network (e.g., Borgatti et al., 2009). This provides a way to examine the influence of leader through which the followership-leadership dynamic is formed in real-life online shaming. The network analysis can also complement the topic modelling analysis that provided insights in Chapter 3 on how online shaming progressed based on what people have expressed. It can be useful to map out the interactions between users and identify whether these interactions occurred within the same group (intragroup) or between different groups (intergroup). Therefore, future studies can use a combination of topic modelling and network analysis to examine a specific case of

online shaming. Future studies could develop a measure of online shaming that captures its complexity and the diverse motives that drive it.

More research remains to be conducted from the perspective of people who had been subjected to shaming, as my thesis focused on examining online shaming from the perspective of people who had engaged in it. Nevertheless, the current findings still have implications for people who are subjected to shaming. Chapter 3 (Study 2) showed that during the online shaming event, groups had not only derogated the outgroup but also supported the ingroup, and the ingroup support had centred on the individuals who were subjected to shaming (i.e., the health minister and the doctor). This suggests that sometimes, people who have been shamed can harness the power by obtaining support from their ingroup. It might help mitigate the negative consequences of online shaming, such as repairing one's reputation. Future studies can continue the research in this area and examine the conditions under which people gather support from the ingroup effectively, and whether the presence of support can influence how individuals frame the shaming experience.

Conclusion

In a commentary piece to the massive online shaming of Justine Sacco (which I discussed in Chapter 4), the researcher and writer Roxane Gay emphasised that social media is like a double-sided sword (Gay, 2013, para. 18): *At its best, social media offers unprecedented opportunities for marginalized people to speak and bring much needed attention to the issues they face. At its worst, social media also offers everyone an unprecedented opportunity to share in collective outrage without reflection.* Indeed, the existing debates on online shaming within and outside of academia, has largely focused on either its "best" or its "worst". This thesis has provided amongst the first empirical evidence that online shaming is a group behaviour driven by intergroup dynamics and the pursuit of shared goals. It is my hope that future research continues down this avenue, to not only

deepen our understanding of why people engage in online shaming, but also inform practical applications that effectively address shaming on social media, whether it is “at its best”, or “at its worst”.

References

References marked with an asterisk indicate studies included in the Scoping Review:

ABC News. (2020, March 7). Melbourne GP clinic closed after doctor tests positive for coronavirus. *ABC News*. https://www.abc.net.au/news/2020-03-07/coronavirus-infects-melbourne-doctor/12023438

*Abraham, B. J. (2014). *Responsible Objects: How human-nonhuman relations reconfigure authority, responsibility, and activism* [PhD Thesis, Western Sydney University]. http://handle.uws.edu.au:8081/1959.7/uws:33169

Abrams, D., Wetherell, M., Cochrane, S., Hogg, M. A., & Turner, J. C. (1990). Knowing what to think by knowing who you are: Self-categorization and the nature of norm formation, conformity and group polarization. *British Journal of Social Psychology*, *29*(2), 97–119. https://doi.org/10.1111/j.2044-8309.1990.tb00892.x

*Adkins, K. (2019). When shaming is shameful: Double standards in online shame backlashes. *Hypatia*, *34*(1), 76–97. https://doi.org/10.1111/hypa.12456

*Aitchison, G., & Meckled-Garcia, S. (2021). Against Online Public Shaming: Ethical Problems with Mass Social Media on JSTOR. *Social Theory and Practice*, *47*(1), 1–31. http://www.jstor.org/stable/45378050

*Amit-Aharon, A., Warshawski, S., & Itzhaki, M. (2023). The role of sense of coherence in workplace violence directed at nurses in the shadow of COVID-19: A cross-sectional study. *Journal of Advanced Nursing*, *79*(12), 4767–4777. https://doi.org/10.1111/jan.15748

Anjum, N., & Katarya, R. (2023). Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities. *International Journal of Information Security*, *23*(1), 577–608. https://doi.org/10.1007/s10207-023-00755-2

*Arancibia, M. C., & Montecino, L. (2017). The construction of anger in comments on the public behavior of members of the social elite in Chile. *Discourse & Society*, *28*(6), 595–613. https://doi.org/10.1177/0957926517721084

Arun, R., Suresh, V., Madhavan, C. E. V., & Murthy, M. N. N. (2010). On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations. In *Lecture notes in computer science* (pp. 391–402). https://doi.org/10.1007/978-3-642-13657-3_43

*Arvanitidis, T. (2016). Publication bans in a Facebook age: How internet vigilantes have challenged the Youth Criminal Justice Act's "Secrecy Laws" following the 2011 Vancouver Stanley Cup riot. *Canadian Graduate Journal of Sociology and Criminology*, *5*(1), 102. https://doi.org/10.15353/cgjsc-rcessc.v5i1.142

Bandura, A. (1999). Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review*, *3*(3), 193–209. https://doi.org/10.1207/s15327957pspr0303_3

*Barron, A. C., Woodyatt, L., Thomas, E. F., Loh, J. E. K., & Dunning, K. (2023). Doing good or feeling good? Justice concerns predict online shaming via deservingness and schadenfreude. *Computers in Human Behavior Reports*, *11*, 100317. https://doi.org/10.1016/j.chbr.2023.100317

Barron, A. T. J., & Bollen, J. (2022). Quantifying collective identity online from self-defining hashtags. *Scientific Reports*, *12*, 15044. https://doi.org/10.1038/s41598-022-19181-w

*Basak, R., Ghosh, S. K., & Sural, S. (2020). Influence modeling of opinion switching by Twitter users in public shaming events. *Social Network Analysis and Mining*, *10*(1). https://doi.org/10.1007/s13278-020-00698-9

*Basak, R., Sural, S., Ganguly, N., & Ghosh, S. K. (2019). Online public shaming on Twitter: Detection, analysis, and mitigation. *IEEE Transactions on Computational Social Systems*, *6*(2), 208–220. https://doi.org/10.1109/tcss.2019.2895734

*Basak, R., Sural, S., & Ghosh, S. K. (2023). Progress tracking and responding to online public shaming events on Twitter. *IEEE Transactions on Computational Social Systems*, *11*(2), 2091–2104. https://doi.org/10.1109/tcss.2023.3295131

BBC. (2020, March 9). Coronavirus: "Shaming" of Australia GP with virus angers doctors. *British Broadcasting Corporation*. https://www.bbc.com/news/world-australia-51796342

Becker, J. C. (2020). Ideology and the promotion of social change. *Current Opinion in Behavioral Sciences*, *34*, 6–11. https://doi.org/10.1016/j.cobeha.2019.10.005

Becker, J. C., & Tausch, N. (2013). When group memberships are negative: The concept, measurement, and behavioral implications of psychological disidentification. *Self and Identity*, *13*(3), 294–321. https://doi.org/10.1080/15298868.2013.819991

*Behera, R. K., Bala, P. K., Rana, N. P., & Kayal, G. (2022). Self-promotion and online shaming during COVID-19: A toxic combination. *International Journal of Information Management Data Insights*, *2*(2), 100117. https://doi.org/10.1016/j.jjimei.2022.100117

Bergquist, M., & Nilsson, A. (2016). I saw the sign: Promoting energy conservation via normative prompts. *Journal of Environmental Psychology*, *46*, 23–31. https://doi.org/10.1016/j.jenvp.2016.03.005

*Bhargava, V. R. (2018). Firm responses to mass outrage: technology, blame, and employment. *Journal of Business Ethics*, *163*(3), 379–400. https://doi.org/10.1007/s10551-018-4043-7

Biddle, S. (2013, December 20). *And Now, a Funny Holiday Joke from IAC's PR Boss*. Gawker. https://web.archive.org/web/20131222094934/https://valleywag.gawker.com/and-now-a-funny-holiday-joke-from-iacs-pr-boss-1487284969

Biddle, S. (2014, December 20). *Justine Sacco is good at her job, and how I came to peace with her*. Gawker. https://web.archive.org/web/20141220162619/http://gawker.com/justine-sacco-is-good-at-her-job-and-how-i-came-to-pea-1653022326

Billig, M. (1996). *Arguing and Thinking: A rhetorical approach to social Psychology* (2nd ed.). Cambridge University Press. https://psycnet.apa.org/record/1996-97418-000

*Billingham, P., & Parr, T. (2020). Enforcing social norms: The morality of public shaming. *European Journal of Philosophy*, *28*(4), 997–1016. https://doi.org/10.1111/ejop.12543

Blackwood, L., & Louis, W. (2017). Choosing between conciliatory and oppositional leaders: The role of out-group signals and in-group leader candidates' collective action tactics. *European Journal of Social Psychology*, *47*(3), 320–336. https://doi.org/10.1002/ejsp.2249

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, *1*(1). https://doi.org/10.1214/07-aoas114

*Blitvich, P. G. (2021). Getting into the mob: A netnographic, case study approach to online public shaming. In M. Johansson, S.-K. Tanskanen, & J. Chovanec (Eds.), *Analyzing Digital Discourses: Between Convergence and Controversy* (pp. 247–274). Springer International Publishing. https://doi.org/10.1007/978-3-030-84602-2_10

*Blitvich, P. G. (2022). Moral emotions, good moral panics, social regulation, and online public shaming. *Language & Communication*, *84*, 61–75. https://doi.org/10.1016/j.langcom.2022.02.002

Bliuc, A., Bouguettaya, A., & Felise, K. D. (2021). Online Intergroup polarization across Political fault lines: An Integrative review. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.641215

Bliuc, A., McGarty, C., Reynolds, K., & Muntele, D. (2007). Opinion-based group membership as a predictor of commitment to political action. *European Journal of Social Psychology*, *37*(1), 19–32. https://doi.org/10.1002/ejsp.334

Bodaghi, A., Fung, B. C. M., & Schmitt, K. A. (2023). Technological Solutions to online toxicity: Potential and pitfalls. *IEEE Technology and Society Magazine*, *42*(4), 57–65. https://doi.org/10.1109/mts.2023.3340235

Borgatti, S. P., Mehra, A., Brass, D. J., & Labianca, G. (2009). Network analysis in the Social Sciences. *Science*, *323*(5916), 892–895. https://doi.org/10.1126/science.1165821

Bosson, J. K., Kuchynka, S. L., Parrott, D. J., Swan, S. C., & Schramm, A. T. (2020). Injunctive norms, sexism, and misogyny network activation among men. *Psychology of Men & Masculinity*, *21*(1), 124–138. https://doi.org/10.1037/men0000217

*Brady, W. J., & Crockett, M. (2019). How effective is online outrage? *Trends in Cognitive Sciences*, *23*(2), 79–80. https://doi.org/10.1016/j.tics.2018.11.004

*Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The MAD Model of Moral Contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, *15*(4), 978–1010. https://doi.org/10.1177/1745691620917336

*Brady, W. J., McLoughlin, K., Doan, T. N., & Crockett, M. J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, *7*(33). https://doi.org/10.1126/sciadv.abe5641

Braithwaite, J. (1989). *Crime, shame and reintegration*. Cambridge University Press.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brehm, J. W. (1966). *A theory of psychological reactance*. Academic Press.

*Bu, Q. (2013). "Human flesh search" in China: the double-edged sword. *International Data Privacy Law*, *3*(3), 181–196. https://doi.org/10.1093/idpl/ipt011

Cabrera, B., Ross, B., Röchert, D., Brünker, F., & Stieglitz, S. (2021). The influence of community structure on opinion expression: an agent-based model. *Journal of Business Economics*, *91*(9), 1331–1355. https://doi.org/10.1007/s11573-021-01064-7

Cambridge University Press. (n.d.). covidiot. In *Cambridge Advanced Learner's Dictionary & Thesaurus*. Retrieved February 18, 2025, from https://dictionary.cambridge.org/dictionary/english/covidiot#google_vignette

*Campano, E. (2020). *Online Shaming : Ethical Tools for Human-Computer Interaction Designers* [Dissertation, Umeå University]. https://umu.diva-portal.org/smash/get/diva2:1447060/FULLTEXT01.pdf

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing*, *72*(7–9), 1775–1781. https://doi.org/10.1016/j.neucom.2008.06.011

Carlsmith, K. M. (2006). The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology*, *42*(4), 437–451. https://doi.org/10.1016/j.jesp.2005.06.007

Carlsmith, K. M., & Darley, J. M. (2008). Psychological aspects of retributive justice. In *Advances in experimental social psychology* (pp. 193–236). https://doi.org/10.1016/s0065-2601(07)00004-4

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish? Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, *83*(2), 284–299. https://doi.org/10.1037/0022-3514.83.2.284

*Chang, L. Y. (2018). Internet vigilantism: Co-production of security and compliance in the digital age. In L. Y. C. Chang & R. Brewer (Eds.), *Criminal Justice and Regulation Revisited* (pp. 147–160). Routledge. https://doi.org/10.4324/9781315174044-9

*Chang, L. Y. C., & Poon, R. (2017). Internet vigilantism: Attitudes and experiences of university students toward cyber crowdsourcing in Hong Kong. *International Journal of Offender Therapy and Comparative Criminology*, *61*(16), 1912–1932. https://doi.org/10.1177/0306624x16639037

*Chang, L. Y. C., & Zhu, J. (2020). Taking justice into their own hands: Predictors of netilantism among cyber citizens in Hong Kong. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.556903

Chapman, A. (2020, March 9). Jenny Mikakos doubles down on criticism of Missy Higgins' coronavirus-stricken doctor dad. *7News*. https://7news.com.au/lifestyle/health-wellbeing/jenny-mikakos-doubles-down-on-criticism-of-missy-higgins-coronavirus-stricken-doctor-dad-c-736683

*Cheung, A. (2014). Revisiting privacy and dignity: Online Shaming in the global E-Village. *Laws*, *3*(2), 301–326. https://doi.org/10.3390/laws3020301

*Chia, S. C. (2019). Crowd-sourcing justice: tracking a decade's news coverage of cyber vigilantism throughout the Greater China region. *Information Communication & Society*, *22*(14), 2045–2062. https://doi.org/10.1080/1369118x.2018.1476573

*Chia, S. C. (2020). Seeking justice on the web: How news media and social norms drive the practice of cyber vigilantism. *Social Science Computer Review*, *38*(6), 655–672. https://doi.org/10.1177/0894439319842190

Chien, C. (2024). Leader Identification and disidentification: constructs, measurements, and nomological network. *Journal of Leadership & Organizational Studies*. https://doi.org/10.1177/15480518241301235

Cialdini, R. B., Reno, R. R., & Kallgren, C. A. (1991). A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology*, *58*(6), 1015–1026. https://doi.org/10.1037/0022-3514.58.6.1015

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587

Condor, S., Tileaga, C., & Billig, M. (2013). Political rhetoric. In L. Huddy, D. O. Sears, & J. S. Levy (Eds.), *Oxford handbook of political psychology* (pp. 262–300). Oxford University Press.

Cooper, F., Dolezal, L., & Rose, A. (2023). *COVID-19 and shame: Political emotions and public health in the UK*. Bloomsbury Academic.

*Corradini, E. (2023). The Dark Threads that Weave the web of Shame: A Network Science-Inspired analysis of body Shaming on Reddit. *Information*, *14*(8), 436. https://doi.org/10.3390/info14080436

*Crockett, M. J. (2017). Moral outrage in the digital age. *Nature Human Behaviour*, *1*(11), 769–771. https://doi.org/10.1038/s41562-017-0213-3

Cutler, K. (2013, March 21). A dongle joke that spiraled way out of control. *TechCrunch*. https://techcrunch.com/2013/03/21/a-dongle-joke-that-spiraled-way-out-of-control/

Darley, J. M., Carlsmith, K. M., & Robinson, P. H. (2000). Incapacitation and just deserts as motives for punishment. *Law And Human Behavior*, *24*(6), 659–683. https://doi.org/10.1023/a:1005552203727

*De Vries, A. (2015). The Use of Social Media for Shaming Strangers: Young People's Views. In *2015 48th Hawaii International Conference on System Sciences* (pp. 2053–2062). https://doi.org/10.1109/hicss.2015.245

Deveaud, R., SanJuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Document Numérique*, *17*(1), 61–84. https://doi.org/10.3166/dn.17.1.61-84

*Dilmaç, J. A. (2014). Looking for the gaze: The case of humiliation in the digital era. *Akademik İncelemeler Dergisi*, *9*(1), 183–203. https://doi.org/10.17550/aid.23284

*Direk, Z. (2020). Politics of Shame in Turkey: Public shaming and mourning. *Sophia*, *59*(1), 39–56. https://doi.org/10.1007/s11841-020-00772-x

Dolezal, L., & Rose, A. (2020, April 7). Naming and shaming: Covid-19 and the medical professional. *Medical Humanities*. https://blogs.bmj.com/medical-humanities/2020/04/07/naming-and-shaming-covid-19-and-the-medical-professional/

Doosje, B., Spears, R., Ellemers, N., & Koomen, W. (1999). Perceived group variability in intergroup relations: the distinctive role of social identity. *European Review of Social Psychology*, *10*(1), 41–74. https://doi.org/10.1080/14792779943000017

*Douglas, D. M. (2016). Doxing: a conceptual analysis. *Ethics and Information Technology*, *18*(3), 199–210. https://doi.org/10.1007/s10676-016-9406-0

Drury, J., & Reicher, S. (1999). The Intergroup Dynamics of Collective Empowerment: Substantiating the social identity model of crowd behavior. *Group Processes & Intergroup Relations*, *2*(4), 381–402. https://doi.org/10.1177/1368430299024005

Drury, J., & Reicher, S. (2005). Explaining enduring empowerment: a comparative study of collective action and psychological outcomes. *European Journal of Social Psychology*, *35*(1), 35–58. https://doi.org/10.1002/ejsp.231

Drury, J., & Reicher, S. (2020). Crowds and collective behavior. *Oxford Research Encyclopedia of Psychology*. https://doi.org/10.1093/acrefore/9780190236557.013.304

*Duncan, S. (2020). Why all the outrage? Viral media as corrupt play shaping mainstream media narratives. *Westminster Papers in Communication and Culture*, *15*(1), 37–52. https://doi.org/10.16997/wpcc.317

*Dunsby, R. M., & Howes, L. M. (2019). The NEW adventures of the digital vigilante! Facebook users' views on online naming and shaming. *Australian & New Zealand Journal of Criminology*, *52*(1), 41–59. https://doi.org/10.1177/0004865818778736

Efron, B., & Hastie, T. (2016). *Computer age statistical inference*. https://doi.org/10.1017/cbo9781316576533

Ellemers, N., Spears, R., & Doosje, B. (2002). Self and social identity. *Annual Review of Psychology*, *53*(1), 161–186. https://doi.org/10.1146/annurev.psych.53.100901.135228

Ellemers, N., & Van Den Bos, K. (2012). Morality in groups: on the Social-Regulatory functions of right and wrong. *Social and Personality Psychology Compass*, *6*(12), 878–889. https://doi.org/10.1111/spc3.12001

Ellemers, N., Van Rijswijk, W., Bruins, J., & De Gilder, D. (1998). Group commitment as a moderator of attributional and behavioural responses to power use. *European Journal of Social Psychology*, *28*(4), 555–573. https://doi.org/10.1002/(sici)1099-0992(199807/08)28:4<555::aid-ejsp879>3.0.co;2-w

Fahey, J. J., Roberts, D. C., & Utych, S. M. (2022). Principled or partisan? The effect of cancel culture framings on support for free speech. *American Politics Research*, *51*(1), 69–75. https://doi.org/10.1177/1532673x221087601

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/brm.41.4.1149

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior*, *25*(2), 63–87. https://doi.org/10.1016/s1090-5138(04)00005-4

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology Research and Practice*, *40*(5), 532–538. https://doi.org/10.1037/a0015808

Finch, W. H., Finch, M. E. H., McIntosh, C. E., & Braun, C. (2018). The use of topic modeling with latent Dirichlet analysis with open-ended survey items. *Translational Issues in Psychological Science*, *4*(4), 403–424. https://doi.org/10.1037/tps0000173

Fitts, A. S. (2017, March 24). Tech's harassment crisis now has an arsenal of smoking guns. *WIRED*. https://www.wired.com/2017/03/techs-harassment-crisis-now-has-an-arsenal-of-smoking-guns

Foster, M. D., Tassone, A., & Matheson, K. (2020). Tweeting about sexism motivates further activism: A social identity perspective. *British Journal of Social Psychology*, *60*(3), 741–764. https://doi.org/10.1111/bjso.12431

Franklin, J. (2013, March 21). A difficult situation. *Twiilo SendGrid*. Retrieved October 20, 2024, from https://sendgrid.com/en-us/blog/a-difficult-situation

Frimer, J. A., & Skitka, L. J. (2018). The Montagu Principle: Incivility decreases politicians' public approval, even with their political base. *Journal of Personality and Social Psychology*, *115*(5), 845–866. https://doi.org/10.1037/pspi0000140

*Frye, H. (2021). The technology of public shaming. *Social Philosophy and Policy*, *38*(2), 128–145. https://doi.org/10.1017/s0265052522000085

Gannes, L. (2013, March 27). *Fired SendGrid developer evangelist Adria Richards speaks out*. AllThingsD. Retrieved October 20, 2024, from https://allthingsd.com/20130327/fired-sendgrid-developer-evangelist-adria-richards-speaks-out/

*Gao, L. (2013). *The Human Flesh Search Engine in China: a case-oriented approach to understanding online collective action* [Doctoral thesis, Loughborough University]. https://core.ac.uk/download/pdf/288381061.pdf

Gay, R. (2013, December 23). Justine Sacco's aftermath: The cost of Twitter outrage. *Salon*. https://www.salon.com/2013/12/23/justine_saccos_aftermath_the_cost_of_twitter_outrage/

*Ge, X. (2020). Social media reduce users' moral sensitivity: Online shaming as a possible consequence. *Aggressive Behavior*, *46*(5), 359–369. https://doi.org/10.1002/ab.21904

*Goldman, L. M. (2015). Trending now: the use of social media websites in public shaming punishments. □*the* □*American Criminal Law Review*, *52*(2), 415. http://www.americancriminallawreview.com/files/4314/2686/4810/TRENDING_NOW_THE_USE_OF_SOCIAL_MEDIA_WEBSITES_IN_PUBLIC_SHAMING_PUNISHMENTS.pdf

Gollwitzer, A., Martel, C., Brady, W. J., Pärnamets, P., Freedman, I. G., Knowles, E. D., & Van Bavel, J. J. (2020). Partisan differences in physical distancing are linked to health outcomes during the COVID-19 pandemic. *Nature Human Behaviour*, *4*(11), 1186–1197. https://doi.org/10.1038/s41562-020-00977-7

Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal*, *26*, 91–108. https://doi.org/10.1111/j.1471-1842.2009.00848.x

greenrd. (2013). *Forking and dongle jokes don't belong at tech conferences [Online forum post]*. Reddit. Retrieved October 20, 2024, from https://www.reddit.com/r/programming/comments/1amhhi/forking_and_dongle_jokes_dont_belong_at_tech/

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl_1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The promise and Pitfalls of Automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

*Gruber, M., Mayer, C., & Einwiller, S. A. (2020). What drives people to participate in online firestorms? *Online Information Review*, *44*(3), 563–581. https://doi.org/10.1108/oir-10-2018-0331

Haslam, S. A., Oakes, P. J., Reynolds, K. J., & Turner, J. C. (1999). Social identity salience and the emergence of stereotype consensus. *Personality and Social Psychology Bulletin*, *25*(7), 809–818. https://doi.org/10.1177/0146167299025007004

Haslam, S. A., & Platow, M. J. (2001). The Link between Leadership and Followership: How Affirming Social Identity Translates Vision into Action. *Personality and Social Psychology Bulletin*, *27*(11), 1469–1479. https://doi.org/10.1177/01461672012711008

Haslam, S. A., & Reicher, S. (2007). Identity Entrepreneurship and the Consequences of Identity Failure: The Dynamics of Leadership in the BBC Prison Study. *Social Psychology Quarterly*, *70*(2), 125–147. https://doi.org/10.1177/019027250707000204

Haslam, S. A., & Reicher, S. D. (2017). 50 years of "Obedience to Authority": From blind conformity to engaged followership. *Annual Review of Law and Social Science*, *13*(1), 59–78. https://doi.org/10.1146/annurev-lawsocsci-110316-113710

Haslam, S. A., Reicher, S. D., & Birney, M. E. (2014). Nothing by Mere Authority: Evidence that in an Experimental Analogue of the Milgram Paradigm Participants are Motivated not by Orders but by Appeals to Science. *Journal of Social Issues*, *70*(3), 473–488. https://doi.org/10.1111/josi.12072

Haslam, S. A., Reicher, S. D., Millard, K., & McDonald, R. (2015). 'Happy to have been of service': The Yale archive as a window into the engaged followership of participants in Milgram's 'obedience' experiments. *British Journal of Social Psychology*, *54*(1), 55–83. https://doi.org/10.1111/bjso.12074

Haslam S. A., Reicher S. D., & Platow M. J. (2010). The New Psychology of Leadership : Identity, Influence, and Power. In *The New Psychology of Leadership* (1st ed., pp. 1–267). Psychology Press. https://doi.org/10.4324/9780203833896

Haslam, S. A., Reicher, S. D., Selvanathan, H. P., Gaffney, A. M., Steffens, N. K., Packer, D., Van Bavel, J. J., Ntontis, E., Neville, F., Vestergren, S., Jurstakova, K., & Platow, M. J. (2023). Examining the role of Donald Trump and his supporters in the 2021 assault on the U.S. Capitol: A dual-agency model of identity leadership and engaged followership. *The Leadership Quarterly*, *34*(2), 101622. https://doi.org/10.1016/j.leaqua.2022.101622

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning. In *Springer series in statistics*. https://doi.org/10.1007/978-0-387-84858-7

*Haugh, M. (2022). (Online) public denunciation, public incivilities and offence. *Language & Communication*, *87*, 44–59. https://doi.org/10.1016/j.langcom.2022.07.002

Hayes, A. F. (2017). *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach* (2nd ed.). Guilford Publications.

Heale, R., & Forbes, D. (2013). Understanding triangulation in research. *Evidence-Based Nursing*, *16*(4), 98. https://doi.org/10.1136/eb-2013-101494

Heng, Y. T., Wagner, D. T., Barnes, C. M., & Guarana, C. L. (2018). Archival research: Expanding the methodological toolkit in social psychology. *Journal of Experimental Social Psychology*, *78*, 14–22. https://doi.org/10.1016/j.jesp.2018.04.012

*Heo, M., & Park, J. (2019). Shame and vicarious shame in the news: A case study of the Sewol ferry disaster. *Journalism*, *20*(12), 1611–1629. https://doi.org/10.1177/1464884916688928

Herbertson, L. (2020, March 8). Missy Higgins' father hits out at "political grandstanding" after testing positive for coronavirus. *SBS News*. https://www.sbs.com.au/news/article/missy-higgins-father-hits-out-at-political-grandstanding-after-testing-positive-for-coronavirus/pj3cw3381

*Hess, K., & Waller, L. (2013). The digital pillory: media shaming of 'ordinary' people for minor crimes. *Continuum*, *28*(1), 101–111. https://doi.org/10.1080/10304312.2013.854868

Hitchick, M. (2020, March 7). "Flabbergasted": Melbourne doctor with coronavirus symptoms continued seeing patients. *The Guardian*. https://www.theguardian.com/world/2020/mar/07/flabbergasted-melbourne-doctor-with-coronavirus-symptoms-continued-seeing-patients

Hogg, M. A., & Reid, S. A. (2006). Social Identity, Self-Categorization, and the communication of group norms. *Communication Theory*, *16*(1), 7–30. https://doi.org/10.1111/j.1468-2885.2006.00003.x

Hogg, M. A., & Smith, J. R. (2007). Attitudes in social context: A social identity perspective. *European Review of Social Psychology*, *18*(1), 89–131. https://doi.org/10.1080/10463280701592070

Hogg, M. A., Turner, J. C., & Davidson, B. (1990). Polarized Norms and Social Frames of Reference: A Test of the Self-Categorization Theory of Group Polarization. *Basic and Applied Social Psychology*, *11*(1), 77–100. https://doi.org/10.1207/s15324834basp1101_6

Hogg, M., & Giles, H. (2012). Norm talk and identity in intergroup communication. In *The Handbook of Intergroup Communication* (pp. 395–410). Routledge. https://doi.org/10.4324/9780203148624-34

*Hou, Y., Jiang, T., & Wang, Q. (2017). Socioeconomic status and online shaming: The mediating role of belief in a just world. *Computers in Human Behavior*, *76*, 19–25. https://doi.org/10.1016/j.chb.2017.07.003

*Huffman, E. M. (2016). *Call-out culture: How online shaming affects social media participation in young adults* [MA thesis, Gonzaga University]. https://www.proquest.com/openview/c82b2ef9569f3c2e886399172e283d53/

Humphreys, J. (2017, December 5). Before you join the online mob, think. You could be next. *The Irish Times*. https://www.irishtimes.com/culture/1.3309605

*Ingraham, C., & Reeves, J. (2016). New media, new panics. *Critical Studies in Media Communication*, *33*(5), 455–467. https://doi.org/10.1080/15295036.2016.1227863

*Jacobs, K., Sandberg, L., & Spierings, N. (2020). Twitter and Facebook: Populists' double-barreled gun? *New Media & Society*, *22*(4), 611–633. https://doi.org/10.1177/1461444819893991

Jalal, W. (2023, September 2). Celebrities are not the only ones affected by "cancel culture". Countries are too — and Singapore wants it gone. *ABC News*. https://www.abc.net.au/news/2023-09-03/singapore-to-outlaw-cancel-culture-is-it-possible/102705826

*Jane, E. A. (2016). Online misogyny and feminist digilantism. *Continuum*, *30*(3), 284–297. https://doi.org/10.1080/10304312.2016.1166560

*Jane, E. A. (2017). Feminist digilante responses to a Slut-Shaming on Facebook. *Social Media + Society*, *3*(2), 205630511770599. https://doi.org/10.1177/2056305117705996

Jigsaw. (2023, April 22). Reducing Toxicity in Large Language Models with Perspective API. *Medium*. Retrieved September 18, 2024, from https://medium.com/jigsaw/reducing-toxicity-in-large-language-models-with-perspective-api-c31c39b7a4d7

*Johnen, M., Jungblut, M., & Ziegele, M. (2018). The digital outcry: What incites participation behavior in an online firestorm? *New Media & Society*, *20*(9), 3140–3160. https://doi.org/10.1177/1461444817741883

Jost, J. T., Becker, J., Osborne, D., & Badaan, V. (2017). Missing in (collective) action: Ideology, system justification, and the motivational antecedents of two types of protest behavior. *Current Directions in Psychological Science*, *26*(2), 99–108. https://doi.org/10.1177/0963721417690633

Kearney, M. W., Sancho, L. R., Wickham, H., Heiss, A., Briatte, F., & Sidi, J. (2022). *Package "rtweet"* [Software].

https://web.archive.org/web/20221125172453/https://cran.r-project.org/web/packages/rtweet/rtweet.pdf

Kim, B. (2020). Effects of social grooming on incivility in COVID-19. *Cyberpsychology Behavior and Social Networking*, *23*(8), 519–525. https://doi.org/10.1089/cyber.2020.0201

Kim, C., & Yang, S. (2017). Like, comment, and share on Facebook: How each behavior differs from the other. *Public Relations Review*, *43*(2), 441–449. https://doi.org/10.1016/j.pubrev.2017.02.006

*Kitchin, P., Paramio-Salcines, J. L., & Walters, G. (2019). Managing organizational reputation in response to a public shaming campaign. *Sport Management Review*, *23*(1), 66–80. https://doi.org/10.1016/j.smr.2019.03.009

*Klonick, K. (2016). Re-Shaming The Debate: Social norms, shame, and regulation in an Internet Age. *Maryland Law Review*, *75*(4). https://doi.org/10.2139/ssrn.2638693

Krim, J. (2005, July 6). Subway fracas escalates into test of the internet's power to shame. *The Washington Post*. Retrieved May 13, 2025, from https://www.washingtonpost.com/archive/business/2005/07/07/subway-fracas-escalates-into-test-of-the-internets-power-to-shame/1759fe23-ef5e-4e29-a850-e1a65190bb5d/

Kubovich, Y. (2015, May 25). Israeli immigration clerk commits suicide after shamed on Facebook for alleged racism. *Haaretz*. Retrieved May 13, 2025, from https://www.haaretz.com/.premium-israeli- clerk-commits-suicide-after-racism-accusations-1.5365606

*Laidlaw, E. B. (2017). Online shaming and the right to privacy. *Laws*, *6*(1), 3. https://doi.org/10.3390/laws6010003

*Larrain, B. (2023). online public shaming as a feminist practice for social change, a double-edged sword for fighting gender violence: the case of the 'feminist funa' in Chile. *Feminist Review*, *135*(1), 80–97. https://doi.org/10.1177/01417789231200487

*Lauricella, S. (2019). Darkness as the frenemy: social media, student shaming, and building academic culture. *Communication Education*, *68*(3), 386–393. https://doi.org/10.1080/03634523.2019.1609055

*Laywine, N. (2021). Selfies or self-development? Humanitarians of Tinder (HoT) and online shaming as a moral community. *First Monday*, *26*(4). https://doi.org/10.5210/fm.v26i4.11673

Leopold, J., Lambert, J. R., Ogunyomi, I. O., & Bell, M. P. (2021). The hashtag heard round the world: how #MeToo did what laws did not. *Equality Diversity and Inclusion an International Journal*, *40*(4), 461–476. https://doi.org/10.1108/edi-04-2019-0129

Levy, B. (2011). Tar and Feathers. *Journal of the Historical Society*, *11*(1), 85–110. https://doi.org/10.1111/j.1540-5923.2010.00323.x

Li, L., Fan, L., Atreja, S., & Hemphill, L. (2024). "HOT" ChatGPT: The promise of ChatGPT in detecting and discriminating hateful, offensive, and toxic comments on social media. *ACM Transactions on the Web*, *18*(2), 1–36. https://doi.org/10.1145/3643829

Locke, E. A., & Latham, G. P. (1990). *A theory of goal setting & task performance*. Prentice-Hall, Inc. https://psycnet.apa.org/record/1990-97846-000

*Loveluck, B. (2019). The many shades of digital vigilantism. A typology of online self-justice. *Global Crime*, *21*(3–4), 213–241. https://doi.org/10.1080/17440572.2019.1614444

Lu, X. (2004). *Rhetoric of the Chinese cultural revolution: The impact on Chinese thought, culture, and communication*. https://doi.org/10.2307/j.ctv10tq3n6

*MacPherson, E., & Kerr, G. (2020). Online public shaming of professional athletes: Gender matters. *Psychology of Sport and Exercise*, *51*, 101782. https://doi.org/10.1016/j.psychsport.2020.101782

Maher, P. J., MacCarron, P., & Quayle, M. (2020). Mapping public health responses with attitude networks: the emergence of opinion-based groups in the UK's early COVID-19 response phase. *British Journal of Social Psychology*, *59*(3), 641–652. https://doi.org/10.1111/bjso.12396

*Mahmood, A., Hashim, H. N. M., Zain, F. M., Suhaimi, N. S., & Yahya, N. A. (2018). A survey on the culture of online shaming: a Malaysian Experience. *International Journal of Academic Research in Business and Social Sciences*, *8*(10), 1125–1134. https://doi.org/10.6007/ijarbss/v8-i10/5270

*Mallén, A. (2016). Stirring up virtual punishment: a case of citizen journalism, authenticity and shaming. *Journal of Scandinavian Studies in Criminology and Crime Prevention*, *17*(1), 3–18. https://doi.org/10.1080/14043858.2016.1157940

*Marwick, A. E. (2021). Morally motivated networked harassment as normative reinforcement. *Social Media + Society*, *7*(2), 205630512110213. https://doi.org/10.1177/20563051211021378

Masullo, G. M. (2022). Facebook reactions as heuristics: Exploring relationships between reactions and commenting frequency on news about COVID-19. *First Monday*. https://doi.org/10.5210/fm.v27i8.12674

Matei, A. (2019, November 1). Call-out culture: how to get it right (and wrong). *The Guardian*. https://www.theguardian.com/lifeandstyle/2019/nov/01/call-out-culture-obama-social-media

Max, D. T. (2020, September 21). The Public-Shaming pandemic. *The New Yorker*. https://www.newyorker.com/magazine/2020/09/28/the-public-shaming-pandemic

McGarty, C., Bliuc, A., Thomas, E. F., & Bongiorno, R. (2009). Collective action as the material expression of opinion-based group membership. *Journal of Social Issues*, *65*(4), 839–857. https://doi.org/10.1111/j.1540-4560.2009.01627.x

McGarty, C., Lala, G., & Douglas, K. M. (2011). Opinion-based groups. In *Cambridge University Press eBooks* (pp. 145–171). https://doi.org/10.1017/cbo9781139042802.008

*Mielczarek, N. (2018). The "Pepper-Spraying Cop" Icon and Its Internet Memes: Social Justice and Public Shaming Through Rhetorical Transformation in Digital Culture. *Visual Communication Quarterly*, *25*(2), 67–81. https://doi.org/10.1080/15551393.2018.1456929

*Milbrandt, T. (2017). Caught on camera, posted online: mediated moralities, visual politics and the case of urban 'drought-shaming.' *Visual Studies*, *32*(1), 3–23. https://doi.org/10.1080/1472586x.2016.1246952

Milgram, S. (1963). Behavioral Study of obedience. *Journal of Abnormal & Social Psychology*, *67*(4), 371–378. https://doi.org/10.1037/h0040525

Milgram, S. (1974). *Obedience to authority: An experimental view*. Tavistock Publications.

Milstein, S. (2013, March 24). I have a few things to say about Adria. *Dogs and Shoes*. https://www.dogsandshoes.com/2013/03/adria.html

*Moore, A. (2016). Shame on You: The role of shame, disgust and humiliation in media representations of 'Gender-Fraud' cases. *Sociological Research Online*, *21*(2), 118–135. https://doi.org/10.5153/sro.3942

*Muir, S. R., Roberts, L. D., Sheridan, L., & Coleman, A. R. (2023). Examining the role of moral, emotional, behavioural, and personality factors in predicting online shaming. *PLoS ONE*, *18*(3), e0279750. https://doi.org/10.1371/journal.pone.0279750

*Muir, S. R., Roberts, L. D., & Sheridan, L. P. (2021). The portrayal of online shaming in contemporary online news media: A media framing analysis. *Computers in Human Behavior Reports*, *3*, 100051. https://doi.org/10.1016/j.chbr.2020.100051

Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, *18*(1). https://doi.org/10.1186/s12874-018-0611-x

*Murumaa-Mengel, M., & Lott, K. (2023). 'Recreational shaming groups of Facebook: Content, rules and modministrators' perspectives'. *Convergence the International Journal of Research Into New Media Technologies*, *29*(4), 944–961. https://doi.org/10.1177/13548565231176184

Nanath, K., & Joy, G. (2021). Leveraging Twitter data to analyze the virality of Covid-19 tweets: a text mining approach. *Behaviour and Information Technology*, *42*(2), 196–214. https://doi.org/10.1080/0144929x.2021.1941259

Nikita, M. (2016). *Package "ldatuning"* [Software]. https://web.archive.org/web/20170718175353/https://cran.r-project.org/web/packages/ldatuning/ldatuning.pdf

*Norlock, K. J. (2017). Online shaming. *Social Philosophy Today*, *33*, 187–197. https://doi.org/10.5840/socphiltoday201762343

Oakes, P. J. (1987). The salience of social categories. In J. C. Turner, M. A. Hogg, P. J. Oakes, S. D. Reicher, & M. S. Wetherell (Eds.), *Rediscovering the social group* (pp. 117–141).

O'Neill, M. (2017, December 11). Gay porn actor Jaxton Wheeler responds to claims he drove August Ames to suicide. *news.com.au*. https://www.news.com.au/finance/work/at-work/gay-porn-actor-jaxton-wheeler-responds-to-claims-he-drove-august-ames-to-suicide/news-story/e5011fb2021fc789991571212ab9dbfb

*Ong, R. (2012). Online vigilante justice Chinese style and privacy in China. *Information & Communications Technology Law*, *21*(2), 127–145. https://doi.org/10.1080/13600834.2012.678653

*Oravec, J. A. (2019). Online social shaming and the moralistic imagination: The emergence of Internet-Based Performative Shaming. *Policy & Internet*, *12*(3), 290–310. https://doi.org/10.1002/poi3.226

Osborne, D., Jost, J. T., Becker, J. C., Badaan, V., & Sibley, C. G. (2019). Protesting to challenge or defend the system? A system justification perspective on collective action. *European Journal of Social Psychology*, *49*(2), 244–269. https://doi.org/10.1002/ejsp.2522

Oz, M., Zheng, P., & Chen, G. M. (2018). Twitter versus Facebook: Comparing incivility, impoliteness, and deliberative attributes. *New Media & Society*, *20*(9), 3400–3419. https://doi.org/10.1177/1461444817749516

*Packiarajah, T. S. (2017). *Online shaming: Exploring factors behind online shaming perpetration as well as its prevalence in adults* [Master thesis, Tilburg University]. https://tilburguniversity.on.worldcat.org/oclc/1362451409

Paez, A. (2017). Gray literature: An important resource in systematic reviews. *Journal of Evidence-Based Medicine*, *10*(3), 233–240. https://doi.org/10.1111/jebm.12266

*Pan, X. (2012). *HUNTED BY THE CROWD: A QUALITATIVE ANALYSIS OF COLLABORATIVE INFORMATION SEARCHING IN CHINA* [PhD dissertation, University of Maryland]. http://hdl.handle.net/1903/14212

*Papp, L. J., Erchull, M. J., Liss, M., Waaland-Kreutzer, L., & Godfrey, H. (2017). Slut-shaming on Facebook: Do social class or clothing affect perceived acceptability? *Gender Issues*, *34*(3), 240–257. https://doi.org/10.1007/s12147-016-9180-7

*Papp, L. J., Hagerman, C., Gnoleba, M. A., Erchull, M. J., Liss, M., Miles-McLean, H., & Robertson, C. M. (2015). Exploring perceptions of slut-shaming on Facebook: Evidence for a reverse sexual double standard. *Gender Issues*, *32*, 57–76. https://doi.org/10.1007/s12147-014-9133-y

Perspective API. (n.d.). *How it works*. Retrieved September 18, 2024, from https://perspectiveapi.com/how-it-works/

Peters, M. D., Marnie, C., Tricco, A. C., Pollock, D., Munn, Z., Alexander, L., McInerney, P., Godfrey, C. M., & Khalil, H. (2020). Updated methodological guidance for the conduct of scoping reviews. *JBI Evidence Synthesis*, *18*(10), 2119–2126. https://doi.org/10.11124/jbies-20-00167

Pew Research Center. (2020). *Republicans, Democrats move even further apart in coronavirus concerns*. Retrieved November 11, 2024, from https://www.pewresearch.org/politics/2020/06/25/republicans-democrats-move-even-further-apart-in-coronavirus-concerns/

Postmes, T., Spears, R., & Lea, M. (2000). The formation of group norms in computer-mediated communication. *Human Communication Research*, *26*(3), 341–371. https://doi.org/10.1111/j.1468-2958.2000.tb00761.x

*Pundak, C., Steinhart, Y., & Goldenberg, J. (2021). Nonmaleficence in Shaming: The ethical dilemma Underlying participation in online public shaming. *Journal of Consumer Psychology*, *31*(3), 478–500. https://doi.org/10.1002/jcpy.1227

*Puryear, C. (2020). *The Threat of Virality: Digital Outrage Combats the Spread of Opposing Ideas* [PhD dissertation, University of South Florida]. https://digitalcommons.usf.edu/cgi/viewcontent.cgi?article=9478&context=etd

*Puryear, C., & Vandello, J. A. (2019). Inflammatory comments elicit less outrage when made in anonymous online contexts. *Social Psychological and Personality Science*, *10*(7), 895–902. https://doi.org/10.1177/1948550618806350

Rafferty, A. C., Hewitt, M. C., Wright, R., Hogarth, F., Coatsworth, N., Ampt, F., Dougall, S., Alpren, C., Causer, L., Coffey, C., Wakefield, A., Campbell, S., Pingault, N., Harlock, M., Smith, K. J., & Kirk, M. D. (2021). COVID-19 in health care workers, Australia 2020. *Communicable Diseases Intelligence*, *45*. https://doi.org/10.33321/cdi.2021.45.57

Rathje, S., Robertson, C., Brady, W. J., & Van Bavel, J. J. (2024). People think that social media platforms do (but should not) amplify divisive content. *Perspectives on Psychological Science*, *19*(5), 781–795. https://doi.org/10.1177/17456916231190392

Record, I., & Miller, B. (2022). People, posts, and platforms: Reducing the spread of online toxicity by contextualizing content and setting norms. *Asian Journal of Philosophy*, *1*(2), 41. https://doi.org/10.1007/s44204-022-00042-2

Reicher, S. (2004). The context of Social Identity: domination, resistance, and change. *Political Psychology*, *25*(6), 921–945. https://doi.org/10.1111/j.1467-9221.2004.00403.x

Reicher, S. D. (1984). The St. Pauls' riot: An explanation of the limits of crowd action in terms of a social identity model. *European Journal of Social Psychology*, *14*(1), 1–21. https://doi.org/10.1002/ejsp.2420140102

Reicher, S. D., Haslam, S. A., & Smith, J. R. (2012). Working toward the experimenter. *Perspectives on Psychological Science*, *7*(4), 315–324. https://doi.org/10.1177/1745691612448482

Reicher, S., Haslam, S. A., & Hopkins, N. (2005). Social identity and the dynamics of leadership: Leaders and followers as collaborative agents in the transformation of social reality. *The Leadership Quarterly*, *16*(4), 547–568. https://doi.org/10.1016/j.leaqua.2005.06.007

Reicher, S., Spears, R., & Postmes, T. (1995). A social identity model of deindividuation phenomena. *European Review of Social Psychology*, *6*(1), 161–198. https://doi.org/10.1080/14792779443000049

Roberts, M. E., Stewart, B. M., & Tingley, D. (2016). Navigating the local modes of big data: the case of topic models. In *Cambridge University Press eBooks* (pp. 51–97). https://doi.org/10.1017/cbo9781316257340.004

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, *91*(2). https://doi.org/10.18637/jss.v091.i02

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., & Rand, D. G. (2014). Structural Topic Models for Open-Ended Survey Responses. *American Journal of Political Science*, *58*(4), 1064–1082. https://doi.org/10.1111/ajps.12103

Ronson, J. (2016). *So you've been publicly shamed* (with a new afterword). Pan Macmillan.

Rösner, L., & Krämer, N. C. (2016). Verbal venting in the social web: Effects of anonymity and group norms on aggressive language use in online comments. *Social Media + Society*, *2*(3). https://doi.org/10.1177/2056305116664220

*Rost, K., Stahel, L., & Frey, B. S. (2016). Digital Social norm Enforcement: Online firestorms in social media. *PLoS ONE*, *11*(6), e0155923. https://doi.org/10.1371/journal.pone.0155923

Russell, N. J. C. (2011). Milgram's obedience to authority experiments: Origins and early evolution. *British Journal of Social Psychology*, *50*(1), 140–162. https://doi.org/10.1348/014466610x492205

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78. https://doi.org/10.1037/0003-066x.55.1.68

Saad, N. (2019, December 19). J.K. Rowling backed a woman who made transphobic remarks. Now she's facing the backlash. *Los Angeles Times*. https://www.latimes.com/entertainment-arts/books/story/2019-12-19/jk-rowling-transphobic-tweet

*Sawaoka, T., & Monin, B. (2018). The paradox of viral outrage. *Psychological Science*, *29*(10), 1665–1678. https://doi.org/10.1177/0956797618780658

*Sawaoka, T., & Monin, B. (2020). Outraged but sympathetic: ambivalent emotions limit the influence of viral outrage. *Social Psychological and Personality Science*, *11*(4), 499–512. https://doi.org/10.1177/1948550619853595

Selah, M. B. (2013, March 25). Black woman techie fired after outing sexist jokesters at conference. *Black Enterprise*. https://www.blackenterprise.com/adria-richards-fired-sexist-jokes-tech-conference/

*Shenton, J. E. (2020). Divided we tweet: The social media poetics of public online shaming. *Cultural Dynamics*, *32*(3), 170–195. https://doi.org/10.1177/0921374020909516

*Šincek, D. (2021). The revised version of the Committing and Experiencing Cyber-Violence Scale and its relation to psychosocial functioning and online behavioral problems. *Societies*, *11*(3), 107. https://doi.org/10.3390/soc11030107

*Skoric, M. M., Chua, J. P. E., Liew, M. A., Wong, K. H., & Yeo, P. J. (2010). Online shaming in the Asian context: community empowerment or civic vigilantism? *Surveillance & Society*, *8*(2), 181–199. https://doi.org/10.24908/ss.v8i2.3485

Smith, L. G. E., Thomas, E. F., Bliuc, A., & McGarty, C. (2024). Polarization is the psychological foundation of collective engagement. *Communications Psychology*, *2*(1). https://doi.org/10.1038/s44271-024-00089-2

Smith, L. G. E., Thomas, E. F., & McGarty, C. (2015). "We must be the change we want to see in the world": Integrating norms and identities through social interaction. *Political Psychology*, *36*(5), 543–557. https://doi.org/10.1111/pops.12180

Spears, R. (2011). Group identities: The social identity perspective. In S. Schwartz, K. Luyckx, & V. Vignoles (Eds.), *Handbook of Identity Theory and Research* (pp. 201–224). Springer. https://doi.org/10.1007/978-1-4419-7988-9_9

Spears, R. (2021). Social influence and group identity. *Annual Review of Psychology*, *72*(1), 367–390. https://doi.org/10.1146/annurev-psych-070620-111818

Spears, R., Lea, M., Corneliussen, R. A., Postmes, T., & Ter Haar, W. (2002). Computer-mediated communication as a channel for social resistance. *Small Group Research*, *33*(5), 555–574. https://doi.org/10.1177/104649602237170

*Spring, V. L., Cameron, C. D., & Cikara, M. (2018). The upside of outrage. *Trends in Cognitive Sciences*, *22*(12), 1067–1069. https://doi.org/10.1016/j.tics.2018.09.006

*Spring, V. L., Cameron, C. D., & Cikara, M. (2019). Asking Different Questions about Outrage: A Reply to Brady and Crockett. *Trends in Cognitive Sciences*, *23*(2), 80–82. https://doi.org/10.1016/j.tics.2018.11.006

Starr, M. (2013, March 22). Two people fired over PyCon "dongle" joke. *CNET*. https://www.cnet.com/news/two-people-fired-over-pycon-dongle-joke/

Steffens, N. K., Haslam, S. A., & Reicher, S. D. (2013). Up close and personal: Evidence that shared social identity is a basis for the 'special' relationship that binds followers to leaders. *The Leadership Quarterly*, *25*(2), 296–313. https://doi.org/10.1016/j.leaqua.2013.08.008

Strauß, S., & Bondü, R. (2021). Links between justice sensitivity and moral reasoning, moral emotions, and moral identity in middle childhood. *Child Development*, *93*(2), 372–387. https://doi.org/10.1111/cdev.13684

*Suhaimi, N. S., Mahmood, A., Yahya, N. A., Zain, F. M., & Hashim, H. N. M. (2018). The Efficacy of Online Shaming as a Modality for Social Control: A Survey amongst UiTM Law Students. *International Journal of Academic Research in Business and Social Sciences*, *8*(12), 903–911. https://doi.org/10.6007/ijarbss/v8-i12/5083

*Sundén, J., & Paasonen, S. (2018). Shameless hags and tolerance whores: feminist resistance and the affective circuits of online hate. *Feminist Media Studies*, *18*(4), 643–656. https://doi.org/10.1080/14680777.2018.1447427

Surani, M., & Mangrulkar, R. (2021). Comparative analysis of deep learning techniques to detect online public shaming. *ITM Web of Conferences*, *40*, 03030. https://doi.org/10.1051/itmconf/20214003030

Swain, S. (2020, March 8). Medical body calls for apology over "flabbergasted" minister's comments about coronavirus GP. *9 News*.

https://www.9news.com.au/national/coronavirus-gp-melbourne-dr-chris-higgins-health-minister/2ec38210-6f2d-427a-ad76-08dced37ebc9

Tait, A. (2020, April 5). Pandemic shaming: Is it helping us keep our distance? *The Guardian*. https://www.theguardian.com/science/2020/apr/04/pandemic-shaming-is-it-helping-us-keep-our-distance

Tajfel, H. (1970). Experiments in intergroup discrimination. *Scientific American*, *223*(5), 96–102. https://doi.org/10.1038/scientificamerican1170-96

Tajfel, H. (1978). *Differentiation between social groups: Studies in the social psychology of intergroup relations.* Academic Press.

Tajfel, H. (1981). *Human groups and social categories: Studies in social psychology*. Cambridge University Press.

Tajfel, H. (1982). Social psychology of intergroup relations. *Annual Review of Psychology*, *33*, 1–39.

Tajfel, H., Billig, M. G., Bundy, R. P., & Flament, C. (1971). Social categorization and intergroup behaviour. *European Journal of Social Psychology*, *1*(2), 149–178. https://doi.org/10.1002/ejsp.2420010202

Tajfel, H., & Turner, J. (1979). An Integrative Theory of Intergroup Conflict. In W. G. Austin & S. Worchel (Eds.), *The social psychology of intergroup relations* (pp. 33–47). Brooks/Cole Pub. Co.

Tajfel, H., & Turner, J. C. (2004). The social identity theory of intergroup behavior. In J. T. Jost & J. Sidanius (Eds.), *Political psychology* (pp. 276–293). Psychology Press. https://www.christosaioannou.com/Tajfel%20and%20Turner%201986.pdf

*Tandoc, E. C., Ru, B. T. H., Huei, G. L., Charlyn, N. M. Q., Chua, R. A., & Goh, Z. H. (2024). #CancelCulture: Examining definitions and motivations. *New Media & Society*, *26*(4), 1944–1962. https://doi.org/10.1177/14614448221077977

Tausch, N., Becker, J. C., Spears, R., Christ, O., Saab, R., Singh, P., & Siddiqui, R. N. (2011). Explaining radical group behavior: Developing emotion and efficacy routes to normative and nonnormative collective action. *Journal of Personality and Social Psychology*, *101*(1), 129–148. https://doi.org/10.1037/a0022728

TED. (2015, July 20). *How one tweet can ruin your life | Jon Ronson* [Video]. YouTube. https://www.youtube.com/watch?v=wAIP6fI0NAI

Terry, D. J., & Hogg, M. A. (1996). Group Norms and the Attitude-Behavior Relationship: A role for Group identification. *Personality and Social Psychology Bulletin*, *22*(8), 776–793. https://doi.org/10.1177/0146167296228002

The salience of social categories. (1987). In *Rediscovering the social group*.

Thomas, E. F., Mavor, K. I., & McGarty, C. (2012). Social identities facilitate and encapsulate action-relevant constructs. *Group Processes & Intergroup Relations*, *15*(1), 75–88. https://doi.org/10.1177/1368430211413619

Thomas, E. F., & McGarty, C. A. (2009). The role of efficacy and moral outrage norms in creating the potential for international development activism through group-based interaction. *British Journal of Social Psychology*, *48*(1), 115–134. https://doi.org/10.1348/014466608x313774

Thomas, E. F., & Osborne, D. (2022). Protesting for stability or change? Definitional and conceptual issues in the study of reactionary, conservative, and progressive collective actions. *European Journal of Social Psychology*, *52*(7), 985–993. https://doi.org/10.1002/ejsp.2912

*Thompson, J. D., & Cover, R. (2022). Digital hostility, internet pile-ons and shaming: A case study. *Convergence*, *28*(6), 1770–1782. https://doi.org/10.1177/13548565211030461

Tjosvold, D., Andrews, I. R., & Struthers, J. T. (1992). Leadership influence: goal interdependence and power. *The Journal of Social Psychology*, *132*(1), 39–50. https://doi.org/10.1080/00224545.1992.9924686

Tollervey, N. (2013, March 27). The Trials, Tribulations and Triumph of PyCon 2013. *The Guardian*. https://www.theguardian.com/info/developer-blog/2013/mar/27/trials-tribulations-triumph-pycon-2013

Traversa, M., Tian, Y., & Wright, S. C. (2023). Cancel culture can be collectively validating for groups experiencing harm. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1181872

*Trottier, D. (2018). Coming to Terms with Shame: Exploring Mediated Visibility against Transgressions. *Surveillance & Society*, *16*(2), 170–182. https://doi.org/10.24908/ss.v16i2.6811

*Trottier, D. (2020a). Denunciation and doxing: towards a conceptual model of digital vigilantism. *Global Crime*, *21*(3–4), 196–212. https://doi.org/10.1080/17440572.2019.1591952

*Trottier, D. (2020b). Confronting the digital mob: Press coverage of online justice seeking. *European Journal of Communication*, *35*(6), 597–612. https://doi.org/10.1177/0267323120928234

Turner, J. C. (1985). Social categorization and the self-concept: A social cognitive theory of group behavior. *Advances in Group Processes: Theory and Research*, *2*, 77–122.

Turner, J. C. (1991). *Social influence*. Thomson Brooks/Cole Publishing Co. https://psycnet.apa.org/record/1992-97487-000

Turner, J. C. (2005). Explaining the nature of power: a three-process theory. *European Journal of Social Psychology*, *35*(1), 1–22. https://doi.org/10.1002/ejsp.244

Turner, J. C. (2010). Towards a cognitive redefinition of the social group. In H. Tajfel (Ed.), *Social Identity and Intergroup Relations* (pp. 15–40). Cambridge University Press. (Original work published 1982)

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group: A self-categorization theory*. Basil Blackwell. https://psycnet.apa.org/record/1987-98657-000

Twardawski, M., & Hilbig, B. E. (2020). The motivational basis of third-party punishment in children. *PLoS ONE*, *15*(11), e0241919. https://doi.org/10.1371/journal.pone.0241919

Van Bavel, J. J., Robertson, C. E., Del Rosario, K., Rasmussen, J., & Rathje, S. (2024). Social Media and Morality. *Annual Review of Psychology*, *75*(1), 311–340. https://doi.org/10.1146/annurev-psych-022123-110258

Van Knippenberg, D., & Hogg, M. (2003). Leadership and Power: identity processes in groups and organizations. In *SAGE Publications Ltd eBooks*. https://doi.org/10.4135/9781446216170

Van Zomeren, M., Postmes, T., & Spears, R. (2008). Toward an integrative social identity model of collective action: A quantitative research synthesis of three socio-psychological perspectives. *Psychological Bulletin*, *134*(4), 504–535. https://doi.org/10.1037/0033-2909.134.4.504

Vogels, E. A. (2022). *A growing share of Americans are familiar with 'cancel culture.'* Pew Research Center. https://www.pewresearch.org/short-reads/2022/06/09/a-growing-share-of-americans-are-familiar-with-cancel-culture/

Vogels, E. A., Anderson, M., Porteus, M., Baronavski, C., Atske, S., McClain, C., Auxier, B., Perrin, A., & Ramshankar, M. (2021). *Americans and 'Cancel culture': Where some see calls for accountability, others see censorship, punishment*. Pew Research Center.

https://www.pewresearch.org/internet/2021/05/19/americans-and-cancel-culture-where-some-see-calls-for-accountability-others-see-censorship-punishment/

Wachs, S., Wright, M. F., & Gámez-Guadix, M. (2024). From hate speech to HateLess. The effectiveness of a prevention program on adolescents' online hate speech involvement. *Computers in Human Behavior*, *157*, 108250. https://doi.org/10.1016/j.chb.2024.108250

*Wall, D. S., & Williams, M. (2007). Policing diversity in the digital age: Maintaining order in virtual communities. *Criminology & Criminal Justice*, *7*(4), 391–415. https://doi.org/10.1177/1748895807082064

Wallace, D. S., Paulson, R. M., Lord, C. G., & Bond, C. F. (2005). Which behaviors do attitudes predict? Meta-Analyzing the effects of social pressure and perceived difficulty. *Review of General Psychology*, *9*(3), 214–227. https://doi.org/10.1037/1089-2680.9.3.214

Waterloo, S. F., Baumgartner, S. E., Peter, J., & Valkenburg, P. M. (2018). Norms of online expressions of emotion: Comparing Facebook, Twitter, Instagram, and WhatsApp. *New Media & Society*, *20*(5), 1813–1831. https://doi.org/10.1177/1461444817707349

*Wehmhoener, K. A. (2010). *Social norm or social harm: An exploratory study of Internet vigilantism* [Master thesis, Iowa State University]. https://doi.org/10.31274/etd-180810-1388

Wenglinsky, M., & Milgram, S. (1975). Obedience to Authority: An Experimental View. *Contemporary Sociology a Journal of Reviews*, *4*(6), 613. https://doi.org/10.2307/2064024

Wenzel, M., & Woodyatt, L. (2025). The power and pitfalls of social norms. *Annual Review of Psychology*. https://doi.org/10.1146/annurev-psych-020124-120310

White, K. M., Smith, J. R., Terry, D. J., Greenslade, J. H., & McKimmie, B. M. (2009). Social influence in the theory of planned behaviour: The role of descriptive, injunctive, and in-group norms. *British Journal of Social Psychology*, *48*(1), 135–158. https://doi.org/10.1348/014466608x295207

Willis, K., Ezer, P., Lewis, S., Bismark, M., & Smallwood, N. (2021). "Covid Just Amplified the Cracks of the System": Working as a Frontline Health Worker during the COVID-19 Pandemic. *International Journal of Environmental Research and Public Health*, *18*(19), 10178. https://doi.org/10.3390/ijerph181910178

Woodley, M. (2020, March 10). Fallout from minister's comments intensifies as GPs stay home. *newsGP*. https://www1.racgp.org.au/newsgp/professional/fallout-from-minister-s-comments-intensifies-as-gp

Yoo, S. W., & De Zúñiga, H. G. (2014). Connecting blog, Twitter and Facebook use with gaps in knowledge and participation. *Communication & Society*, *27*(4), 33–48. https://doi.org/10.15581/003.27.4.33-48

Appendix A: Study 1 Supplementary Materials

Example Initial Search String Used for ProQuest That Contains All Keywords

(noft((cyber OR online OR viral OR digital OR internet OR "social media" OR Twitter OR Facebook OR Instagram OR Weibo) NEAR/3 (shaming OR shame OR ostraci* OR guilt OR humiliat* OR disgust OR contempt OR outrage OR anger OR hate OR immoral* OR vigilantism OR vigilante OR netilantism OR accuse* OR accusation OR condemn* OR "witch hunting" OR "witch-hunting" OR ((community OR group OR social*) NEAR/3 (exclud* OR exclusion OR rejection)))) OR noft("hate mob" OR "internet lynch" OR "cyber lynch" OR "human flesh search" OR "online firework" OR "social media firework" OR doxing OR doxxing))

Example Follow-up Search String Used for ProQuest That Contains All Keywords

noft(("online shaming" OR "public shaming" OR "online firestorm" OR "online shitstorm" OR "crisis shaming" OR "shaming backlash" OR "online outrage" OR "outrage campaign" OR "call out culture" OR "cancel culture" OR "online moral outrage")) AND stype.exact(("Trade Journals" OR "Scholarly Journals" OR "Dissertations & Theses" OR "Reports" OR "Working Papers" OR "Conference Papers & Proceedings" OR "Books") NOT ("Newspapers" OR "Wire Feeds" OR "Blogs, Podcasts, & Websites" OR "Other Sources" OR "Magazines" OR "Audio & Video Works")) AND YR(>=2019)

Table A1

Data Extracted From the Included Articles

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
1	Abraham (2014)[b]	Philosophy	Internet shaming is used as a strategy for activism. Reintegrative shaming can be used as a method to uphold/restore moral order and reintegrate offenders.	This thesis reconceptualised a set of familiar problems, involving the examination of the relationship among authority, responsibility, and activism. The questions raised by online shaming practice and the potential solutions were also discussed.	Other	Abraham (2014) argued that people should acknowledge the existence/responsibility of the "nonhuman face" (i.e., the institutionalised social issues, such as patriarchy and sexism) with a "human face" (i.e., people who were shamed). Both problems are needed to be addressed, yet internet shaming often only targets the "human face" but overlooked the "nonhuman face". Abraham proposed that this problem can be addressed by using reintegrative shaming, in which the "human face" is extracted and acknowledged while criticising the "nonhuman" object itself.	Addressing societal issue
2	Adkins (2019)[a]	Feminist studies	"…I am following Martha Nussbaum's definition of shaming as a stigmatizing judgement, where an actor or group condemns another actor or group for failing to adhere to a shared ideal or norm", wrote Adkins (2018, p. 77).	This research addressed the question that whether online feminist shaming is an effective tool for identifying inappropriate behaviour and effecting social change.	Other	Adkins (2019) suggested that shaming is a risky feminist tactic. It can result in shaming backlashes against the person who initiates shaming, especially when the audience is not sympathetic to the person's judgement of shame.	Addressing societal issue, Consequences, Effectiveness
3	Aitchison & Meckled-Garcia (2021)[a]	Philosophy	Online public shaming was defined as "a form of norm enforcement that involves collectively imposing reputational costs on a person for having a certain kind of moral character" (p. 1), with the aim to disqualify the person from "public discussion and certain normal human relations" (p. 1). Specifically, online shaming has the following key features: 1) characterising an element of someone's moral character as shameful in a more public, massive, and aggregative way than offline shaming; 2) violation of a social norm that the shamers uphold as "moral red lines" (p. 11); 3) serving as an extrajudicial punishment.	This article discusses the ethical problem of online public shaming, including the salient features of social media, key features of online shaming, what is morally wrong with online shaming practice, and lastly, policy recommendations addressing the ethical problems with this practice.	Other	It was argued that online public shaming as an informal punishment fails to meet the due process features, including that the penalties are applied in a transparent and explicit way and through a social-deliberative process, respect people's fundamental rights, being proportionate to the wrongness as well as allowing the punished to participate in the decision-making process (such as defending oneself).	Ethical/legal considerations, Retribution, Value violation

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
4	Amit-Aharon et al. (2023)[a]	Medicine	Online shaming, along with verbal violence and physical violence, is seen as a type of intention to act violently.	This research explored the relationship between sense of coherence, previous exposure to COVID-19, and intention to act violently. Sense of coherence and certain sociodemographic variables were examined as risks or protective factors involved in the intention to act violently.	Self-reported questionnaire	Since online shaming was only included as a vignette representing violence, the findings are not specific to online shaming. It was found, however, a positive correlation between the intentions to engage in verbal violence, online shaming, and physical violence. Verbal violence was also found to be a risk factor for online shaming and physical violence.	Abuse/ Stigmatisation
5	Arancibia & Montecino (2017)[a]	Sociology	"Online denigration ceremonies are often referred to by the term shitstorm……A shitstorm entails the co-participative construction of a negative discourse representation of public social actors engaged in corrupt actions", as defined by Arancibia and Montecino (2017, p. 597)	This research explored the co-participative construction of a shitstorm case.	Critical discourse analysis	The target of the shitstorm (i.e., a businessman) was seen as a representation of Chilean elite members. Profound moral indignation was expressed in the online shitstorm, because of the dissatisfaction towards the corruption, power inequality, and abuse in Chilean society.	Addressing societal issue, Moral outrage
6	Arvanitidis (2016)[a]	Criminology, Sociology	"Naming and shaming" is a form of internet vigilantism that can be defined as the "vigilante justice that occurs in the domain, or with the aid, of the internet". (Arvanitidis, 2016, p. 21)	This research discussed how the Youth Criminal Justice Act was challenged by internet vigilantes.	Other	Through analysing the case study of the 2011 Vancouver riot, the author provided some recommendations for justice officials and social media outlets to modify the vigilante practice and reduce the potential harm caused by such practice.	Ethical/legal considerations, Other social actors
7	Barron et al. (2023)[a]	Psychology	It was mentioned that online shaming where individuals are condemned publicly for violating a norm or value can take the forms of various actions, including sharing images, making social media posts, and leaving comments.	This paper explored whether online shaming is motivated by a justice motive (punish for doing good) or a hedonic motive (punish to feel good, schadenfreude), and whether anonymity and social norms moderate these processes.	Experiments	It was found that people's justice concerns following norm violation increased their online shaming engagement via an increased in the perceived deservingness of the wrongdoer to be called out, followed by an increased schadenfreude about the wrongdoer's suffering. Anonymity was not found to moderate this process and mixed evidence was found for the qualifying effect of social norm.	Value violation, Moral outrage, Entertainment, Retribution
8	Basak et al. (2019)[a]	Computer science	Online public shaming was defined as the act of condemning someone who has violated an accepted social norm, with the aim to elicit the feelings of guilt in the victim.	This article categorised shaming comments collected from Twitter and developed a tool to detect shaming comments based on this categorisation.	Other	Six shaming categories were classified and used for automatic shaming detection: abusive, comparison, passing judgement, religious/ethnic, sarcasm/joke, and "whataboutery". This classification was found to work for a range of shaming events, and the majority of people who commented were likely to shame the victim. The number	Abuse/ Stigmatisation, Value violation

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
						of followers also increased faster for people who contributed shaming comments than those who commented on non-shaming content.	
9	Basak et al. (2020)[a]	Computer science	Online shaming was defined as a mass criticism of a perceived wrongdoer who have violated a social norm.	This study aims to model the polarity of users' opinions during online shaming events and to understand how people change their opinions, based on factors including one's historical opinions, recently held opinions, and others' opinions (i.e., the environment).	Other	It was found that received opinions can change a user more effectively from non-shaming to shaming, than from shaming to non-shaming. However, for transitioning from shaming to non-shaming, the previously sent tweets by the user were more influential. In general, it was found that the users' attributes (e.g., number of followers, number of followers) were more influential in changing their opinions than the opinions of their followers, suggesting that "having greater twitter popularity may not necessarily translate into influence in changing the opinion of a follower" (p. 84).	Social approval/ recognition, Value violation
10	Basak et al. (2023)[a]	Computer science	Online shaming events were described as negative viral events that can cause devastating consequences. The victims of online shaming were often accused of violating a social norm.	This study examines whether victims of online shaming can predict the progress of the event as well as how the victims should respond (denial or apology) to mitigate the progress of online shaming.	Other	It was suggested that compared to a public figure or an organisation, ordinary people who experience online shaming often have limited resources to reduce the impact of online shaming. Using machine learning models in predicting the progress of different shaming events, it was found that the best response should be based on the event progress (i.e., timing and the type of strategy). Specifically, it was recommended that admitting the fault can be considered as an early response, whereas denial might not be suitable when approaching the end stage of online shaming.	Consequences, Value violation
11	Behera et al. (2022)[a]	Computer science	Online shaming was understood as a shaming practice via social media to call-out transgressions. Particularly, it was observed during COVID-19 lockdowns that people shame those who violated the guidelines.	This study applies the black swan theory (i.e., online shaming can be seen as a black swan event) to examine whether the "toxic combination of online shaming and self-promotion" on social media predicts changes in rule-breaking behaviour.	Survey	It was found that relationship building, perceived enjoyment of using social media, and self-presentation positively predicted self-promotion, which led to higher victimisation of online shaming, which in turn, led to greater change in one's self-reported behaviour (i.e., reduction in rule-breaking behaviour). Perceived risk of COVID-19 was found to moderate the relationship between relationship building, perceived enjoyment, and self-presentation with self-promotion. Change in behaviour was found to be higher for females than for	Consequences, Value violation

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
						males, and higher for adults over middle-aged than younger aged adults.	
12	Bhargava (2018)[a]	Business	Online shaming as a mass social media outrage.	This research discussed the ethics of firms or managers' response to fire an employee who faced mass social media outrage.	Other	Bhargava (2018) suggested that firms or managers should not use the practice of firing employees who are involved in mass social media outrage, as firing in this context constitutes an inappropriate, unjustified act of blame.	Ethical/legal considerations, Other social actors
13	Billingham & Parr (2019)[a]	Philosophy	The authors conceptualised public shaming as a sanction imposed on norm violators.	This paper proposed a framework to assess the justifiability of public shaming as a sanction.	Other	Five constraints of public shaming were identified: proportionality, necessity, respect for privacy, non-abusiveness, and reintegration. It was contended that a shaming instance is justifiable if and only if each of these constraints is met. However, most instances of online public shaming failed to meet these constraints.	Abuse/ Stigmatisation, Ethical/legal considerations, Justifiability, Retribution, Value violation
14	Blitvich (2021)[d]	Media and communication	Online public shaming was understood as a concept that is related to smart mob (e.g., human flesh search) and vigilantism.	The author analysed one case of online shaming and compared its defining characteristics with the defining features or phases proposed by previous research on human flesh search and smart mob.	Netnographic analysis, Case study	Through analysing a specific case of online shaming on racism (the victim of shaming was referred to as AS), it was found that inconsistent with the previous research on smart mob and vigilantism, the incident of online shaming seems to not have a clear end, but was referenced when other shaming events denouncing racism happened. Rather, AS's case is not in isolation, but a part of an ongoing denunciation of racism in the USA. It was also found that within the same smart mob, there were dissenting voices, suggesting self-reflexivity was not completely absent in such punitive actions.	Addressing societal issue, Retribution
15	Blitvich (2022)[a]	Language, Media and communication	Online public shaming was understood as fundamentally aggressive. It was also understood as relating to digital vigilantism and smart mob (e.g., digilante, human flesh search engine).	This research examined people's motivations to engage in online shaming and the goals being pursued by digilantes.	Digital discourse analysis/ thematic analysis	Through closely analysing 6 recent cases of online public shaming in the USA, the results suggested that online public shaming can be conceptualised as an offensive behaviour along the lines of batteries moral emotions (moral indignation/outrage, empathy/concern directed at the deviants and the victims), good moral panics (storing the moral imbalance), and social regulation (e.g., through exposing the actions of the deviants and deterring the deviants). Specifically, the motives of the online shaming cases include show emotions, denounce racism, show support for deviant, denounce the endangerment of a child, and denounce the wasting of police time. The goals include	Abuse/ Stigmatisation, Addressing societal issue, Moral outrage, Value violation, Social justice, Retribution, Social/behavioural control

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
						exposing the deviant, holding the deviant accountable, and doxing.	
16	Brady & Crockett (2019)[f]	Psychology	Online moral outrage	This paper discussed the effectiveness of online outrage.	Other	Brady and Crockett argued that although moral outrage can have positive social consequences (e.g., catalysing collective action), it has more downsides than upsides, such as being oppressive and reducing the effectiveness of collective action.	Addressing societal issue, Consequences, Effectiveness, Moral outrage, Justifiability
17	Brady et al. (2020)[a]	Psychology	Online shaming or online firestorm is a type of moral contagion that is comprised of the expression of moral emotions online and the contagion process (i.e., how the information diffuses online).	This paper reviewed how the spread of moralised content online (i.e., moral contagion) is influenced by people's motivation, attention, as well as the design of online platforms.	Other	It was proposed that people's expressing and sharing of moralised content online could be shaped by the motives that are based on group identity. The design of social media platforms might further amplify some of the group-based motives or desires.	Moral outrage, Social belongingness, Social recognition/ approval
18	Brady et al. (2021)[a]	Psychology, Computer science	Moral outrage was defined as including the feelings (composed of moral emotions such as anger) triggered by the perceived violation of one's morals, which could be associated with motives such as "blaming people/events/things, holding them responsible, or wanting to punish them" (p. 2).	This paper examined how social learning (including reinforcement learning via social feedback and norm learning) influenced people's outrage expression on social media.	Observational studies on Twitter, Experiments	It was found that people's future outrage expression was predicted by social feedback specific to expression of moral outrage (e.g., likes and absence of likes), as well as the expressive norms in the networks. This suggests that both reinforcement learning and norms with the social networks/communities affect people's expression of outrage. Though it was also found that in communities where expression of outrage is more common, users were less sensitive to social feedback, suggesting norm learning overshadows the influence of reinforcement learning.	Moral outrage, Social recognition/ approval, Value violation
19	Bu (2013)[a]	Law	Human flesh search (HFS) involves cyber-vigilantes searching and exposing details of a subject who is perceived to have done evil act. It was suggested to have been used for "social shaming, monitoring others, and ostracizing subjects." (p. 182)	This paper provided a legal discussion about the upsides and downsides of human flesh search in China.	Other	Through analysing the case study of Wang v Daqi.com and Zhang, the author suggested that although human flesh search has the pro of Social justice, it also has the con of invading people's privacy. Bu further discussed the administrative and legislative regulations proposed by the Chinese government and suggested that one needs to consider the balance between the rights of one's freedom of speech and privacy.	Consequences, Ethical/legal considerations, Social justice
20	Campano (2020)[b]	Computer Science, Philosophy	Shaming was referred to as the instance where a group of users collectively post on social media to punish someone for "doing something they perceive as unjust, for the ostensible purpose of societal modification" (p. 3). Campano	This study provided ways that enable Human-Computer Interaction designers to develop systems that	Other	Through discussing the different philosopher's views of online shaming as well as the arguments against and for online shaming, Campano proposed three possible solutions: 1) build an online justice system 2) develop algorithms to detect online	Ethical/legal considerations, Other social actors, Retribution, Social justice

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
			also suggested that the definition of online shaming is similar to the definition of online vigilantism, though the former emphasises the punishing nature, whereas the latter focuses on the extrajudicially nature.	ethically address online shaming.		shaming, and 3) build a shaming circuit breaker.	
21	Chang (2018)[d]	Criminology	Internet vigilantism was defined as the behaviour to share and expose information that could help solve crimes or identify the details of wrongdoers who might engaged in corruption, rule-breaking, or other deviating behaviours.	This chapter reviewed the phenomenon of internet vigilantism, including its types, characteristics, ethical/legal concerns, and consequences.	Other	The author suggested that fun-seeking is not the only motivation for people to engage in online vigilantism. Rather, people who engage in online vigilantism perceived it as an effective way to achieve justice and perceived a higher self-efficacy. Nonetheless, this subjective, sense of justice might not warrant real justice as online vigilantism might involve making mistakes, causing collateral damage, violating privacy, and interfering with or even damaging the legal system.	Consequences, Effectiveness, Ethical/legal considerations, Social justice
22	Chang & Poon (2017)[a]	Criminology	Cyber crowdsourcing, or human flesh search, was defined by Chang and Poon (2017) as "informal community guards to" (p. 1914) share and expose information that could help solve crimes or identify the details of wrongdoers who might engaged in criminal or immoral behaviours. The author recognised that it could entail an aim to shame or punish for justice, and fun/curiosity seeking.	This research examined how people perceive internet vigilantism in general and how people with different roles in internet vigilantism (i.e., the experiences of being a victim, a vigilante, and/or a bystander) differ in their internet vigilantism-related perceptions.	Survey	People who have the experiences of engaging in internet vigilantism perceived the criminal justice in Hong Kong as more ineffective compared to the victims of online shaming. People who engaged in internet vigilantism also perceived a higher level of self-efficacy and perceived the practice as more effective in achieving social justice than those who did not engage in online shaming.	Effectiveness, Retribution, Social justice
23	Chang & Zhu (2020)[a]	Psychology	Human flesh search, or netilantism, was defined as a behaviour that involves users to act collectively and coordinately, which can achieve justice via exposing information about individuals/groups.	This research examined whether and how netizens' intention to engage in human flesh search is influenced by their personal characteristics of netizens (gender, time spent online) and their attitudes toward social justice, fairness and criminal justice systems.	Survey	Through analysing respondents' responses to the survey questionnaire ($N = 971$), it was found that people who had less confidence in the fairness of the criminal justice system, believed more in social justice (though this relationship is weak) and vigilantism, showed a more positive attitude towards human flesh search, which led to a greater intention to engage human flesh search.	Retribution, Social justice
24	Cheung (2014)[a]	Law	Cheung suggested that online shaming is a unique form of shaming that acts as a social sanction.	This paper discussed the role of privacy and dignity and their relation to online shaming.	Other	Cheung argued that "the recognition and protection of the dignity and privacy of an individual is necessary in order to arrive at	Ethical/legal considerations, Value violation, Retribution

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
						norms and values inherent in decent participation in the e-village". (p. 301)	
25	Chia (2019a)[a]	Media and Communication	Shaming was described as an aim/purpose of cyber vigilantism.	This research examined the public perception of cyber vigilantism in three societies, China, Taiwan, and Hongkong via analysing the published news reports.	Content analysis and statistical analysis	Most of the collected news stories reported cases of cyber vigilantism. The author found that cyber vigilantism was reported as an effective practice to disclose transgressions and the wrongdoer's identity as well as to reinforce social norms and laws. However, the downside of cyber vigilantism received less media attention. Among three societies, the news coverage in China tends to be more balanced than the news coverage in Hong Kong and Taiwan, as the news in China tends to be longer, and the pros and cons of online vigilantism with both supportive and critical responses were discussed. For example, the news coverage in the greater China region tends to highlight the role of online vigilantism to reinforce norms and laws effectively.	Effectiveness, Justifiability, Other social actors, Social/behavioural control
26	Chia (2019b)[a]	Media and communication	Online shaming was described according to cyber vigilantism, digital vigilantism, or web sleuthing.	This research examined whether and how the exposure to media coverage of cyber vigilantism affects people's evaluations and responses to cyber vigilantism.	Online survey	People's evaluation of cyber vigilantism (i.e., the usefulness, harmfulness, and perceived social approval of the practice) and behavioural intention were aligned with the frequency of media reports and how the practice was framed by the media. It was also found that people's desire for justice and their confidence in police increased their intention to practice cyber vigilantism. Perceived social approval was also found to increase people's behavioural intention indirectly via perceived usefulness and to decrease behavioural intention via perceived harmfulness.	Other social actors, Social justice, Social recognition/ approval
27	Corradini (2023)[a]	Computer science, Media and communication	The focus was on body shaming, a specific subtype of shaming that targets one's body shape and/or appearance.	This research examined how individuals and communities interact and together influence body shaming on Reddit.	Social network analysis, Topic modelling	It was found that across different subreddit (different communities), there were significant differences in the sentiment, number and frequency of comments, length of comments, topics focused on each community, suggesting different subreddits have different norms on commenting behaviours. For example, 1) some provided a more supportive environment with more positive sentiment, whereas some provided a less supportive environment; 2) some encouraged longer comments with more	Abuse/ Stigmatisation

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
						detailed discussion, whereas some encouraged shorter comments; 3) different topics were focused, with some focused more on physical appearance and lifestyle critiques, while others focused more on weight less and exercise.	
28	Crockett (2017)[f]	Psychology	Crockett defined moral outrage as "a powerful emotion that motivates people to shame and punish wrongdoers (Salerno & Peter-Hagene, 2013)" (p. 1).	This article discussed the role of online social media in changing the expression of moral outrage and the social consequences that it can bring.	Other	Crockett suggested that digital media (especially its affordances) might transform moral outrage in terms of the stimuli that trigger outrage, elicited responses, and outcomes (e.g., bringing reputational benefits for the individuals and deepening social divides).	Consequences, Value violation, Moral outrage, Social recognition/ approval
29	de Vries (2015)[c]	Computer Science, Media and communication	de Vries (2015) suggested that there lacks an agree definition on online shaming.	This research examined young people's views on the occurrence and acceptance of online shaming towards strangers.	Focused group	Participants viewed online shaming as justified because of its positive functions and benefits (e.g., deterring similar behaviours). Thus, public online shaming is viewed as different from behaviours such as online bullying.	Justifiability, Social/behavioural control, Value violation
30	Dilmaç (2014)[a]	Sociology	Cyberhumiliation (or cyberbullying, cyberharassment, cyberintimidation) was defined as the behaviour of tarnishing someone's reputation via creating a false profile of someone else against their will, revealing their personal information and images, harassing them via mockery, insults, and threats, and/or having one's profile erased.	This research discussed the relationship between self-exhibition online and cyberhumiliation.	Other	Cyberhumiliation was suggested to be a consequence of individuals overexposing themselves in the digital world. Thus, Dilmac suggested that researchers should focus on 'constant individuals' desire of being "viewable" to everyone in the digital world' (p. 199).	Abuse/ Stigmatisation
31	Direk (2020)[a]	Philosophy, Sociology	Direk (2020) identified two types of online public shaming: 1) "public shaming as an activist strategy of moral reform" (p. 39) and 2) public mourning and shaming as a way to express injustice and resistance against authoritarianism.	This article compared two types of online public shaming, i.e., shaming as a feminist strategy versus public mourning and shaming.	Other	Direk (2020) argued that the first kind of shaming (i.e., shaming as an activist strategy) is "repressive and unfair attacks that destroy public deliberation" (p. 39), while the second type of shaming (i.e., public mourning and shaming) is acceptable as it challenges the injustice and resists against authoritarianism.	Consequences, Justifiability, Social justice
32	Douglas (2016)[a]	Philosophy	Doxing was defined as third party's act of exposing a target's personal information, often "with the intent to humiliate, threaten, intimidate, or punish the identified individual" (Douglas, 2016, p. 199). Three types of doxing were identified and defined: 1) deanonymising doxing (i.e., the identity of the formerly anonymous target is	This research provided a conceptual analysis of doxing behaviour.	Other	Douglas argued that doxing is justified when it only reveals the information about the wrongdoing itself and such revelation is in the public interest, whereas doxing that involves disclosing additional information that allows the target to be identified, harassed, and physically threatened are unjustified.	Abuse/ Stigmatisation, Consequences, Retribution, Justifiability

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
			revealed) 2) targeting doxing (i.e., specific, personal information about the target is disclosed that can lead to offline harassment) and 3) delegitimizing doxing, which attempts to shame, humiliate, and damage the credibility of someone via revealing that individual's intimate personal information.				
33	Duncan (2020)[a]	Media and communication, Sociology	No definition of shaming was provided.	This study examined a case study to understand the influences of negative viral commentary, including how it influenced the emotions expressed on social media (i.e., anger and outrage), traditional media's coverage of the incident, and the response from the organisation who was involved in the incident.	Other	Duncan (2020) argued that viral sports-related commentary is an example of "corrupted play", which stimulates chaos (i.e., moral emotions such as anger and outrage), social divides, as well as makes the news coverage of mainstream more extreme.	Consequences, Moral outrage, Other social actors
34	Dunsby & Howes (2019)[a]	Criminology	Dunsby and Howes defined naming and shaming as a form of digital vigilantism.	This study examined 1) the view of Australian Facebook users regarding naming and shaming people who were suspected or convicted of a crime or violated a norm and 2) people's engagement in the naming and shaming practice.	Online survey	Most participants reported that they had not been involved in naming and shaming of personal suspected or convicted of a crime. Participants perceived this practice as appropriate to warn others, foster awareness, and maintain community wellness. Some believed that this practice reflected acknowledging and supporting others on Facebook. However, the circumstances for appropriate shaming are nuanced: Participants were concerned with circumstances such as the severity of the crime, whether it is suspected or convicted, and whether it impedes justice.	Addressing societal issue, Ethical/legal considerations, Justifiability, Social recognition/ approval
35	Frye (2022)[a]	Philosophy, Political science	The author argued that online shaming can be described as a punishment, as well as a problem of social technology. Specifically, public shaming in general (not specific to online forms) was understood as "a piece of social technology that helps groups achieve particular ends" (p. 130). Through online shaming, "groups express their	The aim of this essay is to argue that online shaming is a problem of social technology.	Other	Public shaming was argued to be a type of social technology, which "advances the ends of a particular group" (p. 135) following a norm violation. It can facilitate public cooperation when shaming is reintegrative. However, according to the author, disintegrative shaming that ostracise individuals from a community does not induce cooperative behaviour. Because	Abuse/ Stigmatisation, Ethical/legal considerations, Retribution, Value violation

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
			value judgments, ostracize individuals, develop solidarity, deter would-be wrongdoers, among other things" (p. 130).			online shaming via social media is in a virtually permanent status and that people who engage in shaming are people who are strangers to the target of shaming (hence no reconciliation can be achieved), online public shaming is a disintegrative shaming that is problematic.	
36	Gao (2013)[b]	Media and communication, Political science	Human flesh search (HFSE) was defined as "a form of online collective action in which more than one Internet user contributed collectively to a certain goal but in different ways" (Gao, 2013, p. 175). HFSE cases were classified into three types based on the different goals: punishing a target, fact-checking credibility, and looking for a missing person. Gao suggested that shaming is essential to HFSE that punishes a target. It also differentiate target-punishing HFSE from other types of HFSE.	This thesis examined the political focus of the phenomenon of human flesh search engine.	Case-oriented approach	Through comparing HFSE with other types of collective actions, it was found that "the internal process of politically-focused HFSE differs largely from that of recent Chinese offline popular protests, which indicates that HFSE does not have an offline equivalent, although some of its stages can be witnessed offline" (p. iii). Specifically, it was also suggested that when HFSE focuses on government/officials, it entails a motive of seeking justice.	Addressing societal issue, Retribution, Social justice
37	Ge (2020)[a]	Psychology	Online shaming was defined as a type of cyberviolence or immoral behaviour.	This research examined whether people's use of social media would influence their moral decision-making process and thus, influence people's tendency to engage in online shaming.	Experiment	Ge (2020) found that social media exposure decreased people's moral intensity (including the awareness of the potential consequences and the social disapproval of shaming practice), which decreased their moral sensitivity. Hence, people were more likely to engage in online shaming.	Abuse/ Stigmatisation, Social recognition/ approval
38	Goldman (2015)[c]	Law	Online shaming is a punishment.	This review discussed the history of public shaming as a criminal punishment, people's reactions to shaming punishment in the modern world, as well as the effectiveness of incorporating online shaming punishment via social media into the judicial system.	Other (review)	Goldman suggested that online public shaming can be justified by multiple justice theories, including deterrence and rehabilitation theories, as well as incapacitation theory. Goldman also suggested that online public shaming punishments can be effective, especially when certain guidelines are followed. For example, avoiding sentencing inconsistency and avoiding punishments for humiliation only. It was suggested that online public shaming punishment that follows these guidelines can be considered to include in the current judicial system.	Effectiveness, Ethical/legal considerations, Retribution, Justifiability
39	Gruber et al. (2020)[a]	Media and communication	Online firestorm can be understood as a communicative action to achieve a goal	This research examined what motivates or constrains people to	Online survey	Collective identity was found to be the strongest positive predictor of online firestorm participation. Involvement	Addressing societal issue

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
			(i.e., solving societal problems and creating social change).	participate in an online firestorm.		recognition (i.e., personal relevance of the problem issue involved in the online firestorm) was found to be the second strongest, positive predictor. Accordingly, anonymity was found to have a hindering effect on people's willingness to participate in online firestorms.	
40	Haugh (2022)[a]	Language, Media and communication	Shaming, or public denunciation, was understood as a form of status degradation ceremony with the aim to stigmatise (rather than to reintegrate) an individual who are denounced.	This paper examines how public incivilities (i.e., conducts in encounters between strangers in the online context) are rendered as offensive on social media.	Interactional pragmatics	Through analysing 26 instances of online public denunciation that are reported in mainstream media outlets, Haugh suggested that posting public incivilities involve active framing of the conduct as a transgression that warrants condemnation by others. Specifically, the denunciations arise from conflicts between strangers in the first place, and are often contested. They involve two key subjectivities: "what is considered 'noteworthy' by the poster in question, and what part of that public encounter is selected as the focal point of condemnation is grounded in the moral world of the denouncer in question" (p. 57).	Abuse/ Stigmatisation, Justifiability, Moral outrage, Other social actors, Value violation
41	Heo & Park (2019)[a]	Media and communication	According to Heo and Park (2019), online shaming can be understood as a punishment for a norm violation, and shaming gives people a sense of shame.	Through examining the news stories on the case of the Sewol ferry disaster, this study examined how some South Koreans felt the shame of others' wrongdoings (i.e., vicarious shame) and how the shaming of those who have violated norms was expressed in the South Korean traditional and online news.	Content analysis	Both shaming and vicarious shame (or group-based shame) were found in the news reports, with the expressions being more frequent via internet than traditional media. Internet media expressed shaming and vicarious shame more frequently than traditional media. Shaming was found to be most frequently appear with anger. Shaming was also found to be more likely to appear when the wrongdoings were confirmed, described in detail, the negative influence was mentioned, or punishment for the wrongdoings was expected.	Moral outrage, Other social actors, Retribution, Value violation
42	Hess & Waller, (2014)[a]	Media and Communication	Digital shaming is a continuum of the traditional form of public shaming.	This article discusses the role media plays in calling out and shaming alongside formal, judiciary punishments imposed on 'ordinary' people.	Other	In history, public shaming has been used as a formal punishment across many societies. Although shaming as a formal punishment phased out during the early 19th century, it remains a powerful practice that extended to social media platforms.	Other social actors, Retribution

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
43	Hou et al. (2017)[a]	Psychology	According to Hou et al. (2017), shaming is defined as a process that can "draw attention to the bad dispositions or actions of an offender", which reflects people's desires to enforce social norms and exercise social control.	This study examined how individual factors and justice beliefs contribute to people's engagement in online shaming.	Survey	It was found that people's socio-economic status and beliefs in a just world influenced people's engagement in online shaming.	Retribution, Value violation, Social justice, Social/behavioural control
44	Huffman (2016)[b]	Media and communication	Huffman (2016) suggested that online shaming is a violence act that occurs to someone who is perceived to have "transgressed social or moral boundaries" (p. 12). It entails an aim to humiliate or punish the perceived wrongdoer, "conjure deep emotional feelings about one's selfworth and modify one's understanding of their place in the world" (p.11).	This study examined how online shaming experience might influence people's participation on social media.	Open-ended survey	Some respondents reduced their posting rates on social media after experiencing or witnessing shaming, due to the anxiety of being verbally attacked by an anonymous crowd.	Consequences
45	Ingraham & Reeves (2016)[a]	Media and communication	Online shaming is a manifestation of contemporary moral panics. Online shaming is used to punish and ostracise moral offenders in a way that the formal justice system often cannot.	This essay examined the role of online shaming in relation to moral panic.	Other	It was argued that expressing moralised content online (i.e., punishing and ostracising others via online shaming) provided people an opportunity to escape from the everyday "powerlessness" and helps them to acquire a sense of "doing something" or at least "making oneself heard". In this sense, online shaming is meaningful political participation. However, the authors questioned whether this temporary distraction from a larger crisis can have a long, lasting effect.	Consequences, Effectiveness, Justifiability, Other social actors, Retribution
46	Jacobs et al. (2020)[a]	Media and communication	Shaming was used by populist politicians to attack and bully journalists (i.e., naming and shaming) and/or to engage followers and reduce the credibility of the press in general (i.e., shaming without naming).	This article examined how social media platforms (i.e., Facebook and Twitter) can be used by populist politicians to engage the public. Tweets and Facebook posts made by Members of Parliament (MPs) of Austria, The Netherlands, and Sweden were examined.	Content analysis and statistical analysis	Among tweets that named a media account, populist MPs engaged in shaming (10.6%) more often than non-populist MPs (3.1%). It was also found that posts made by populist MPs (4.14%) attracted more angry reactions than non-populist MPs (1.09%).	Other social actors, Moral outrage
47	Jane (2016)[a]	Feminist studies, Media and communication	Naming and shaming, or public shaming, is used as an approach to online feminist vigilantism.	This article discussed the ethical concerns and risks of using the feminist vigilante approach as a response	Other	Acts of feminist online vigilantism tend to be ethically questionable and can have uncertain or negative outcomes (such as putting activists at risk and strengthening extrajudicial cultures online).	Addressing societal issue, Consequences, Ethical/legal considerations

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
				to online gendered hate speech.			
48	Jane (2017)[a]	Feminist studies, Media and communication	Naming and shaming, or public shaming, is used as an approach to online feminist vigilantism.	This article examined the feminist vigilante response towards an Australian woman who was "slut-shamed" on Facebook.	Other	Through analysing the case, it was shown that the feminist vigilante response was effective and ethically justified given the lack of interventions at an institutional level. Nonetheless, this practice is often associated with risks and is ethically questionable. Thus, it was suggested that a multifaceted intervention is needed to address online gendered hate systematically.	Addressing societal issue, Effectiveness, Ethical/legal considerations, Justifiability
49	Johnen et al. (2018)[a]	Media and communication	Online firestorm was defined as a crowd-based outrage that targets brands, public institutions, public figures, or other individuals. It often involves negative opinions towards the target and intense indignation. It is also a specific form of moral panic.	This research examined why people engage in online firestorms.	Online experiment and content analysis	It was found that people's perception of the number of people participating in an online firestorm influenced their engagement in the online firestorm, as the more participants engaged in online shaming, the less willing people were to participate because it became more difficult for them to stand out. Hence, it was suggested that the desire for social recognition can be a key driver of online firestorm behaviour.	Social recognition/ approval, Moral outrage
50	Kitchin et al. (2020)[a]	Sport	Online shaming was defined as a new form of public shaming. It involves aims to humiliate or punish someone or some organisation who is perceived to have violated social norms.	This article examined a case of shaming campaign against an English Premier League football club for disability discrimination. Specifically, the authors examined how this campaign was seeking to increase people's awareness, the outcomes of the campaign, as well as the club's response to this campaign.	Content analysis and semi-structured interview	Three categories of tweets were identified: 1) tweets with the purpose to increase awareness 2) tweets criticised the organisation/a specific individual 3) tweets with the purpose to discuss in the broader context of discrimination. The campaign successfully raised people's awareness about the issue (i.e., disability and discrimination) internationally.	Abuse/ Stigmatisation, Addressing societal issue, Consequences, Retribution, Value violation
51	Klonick (2016)[a]	Law	According to Klonick, "online shaming is (1) an over-determined punishment with indeterminate social meaning; (2) not a calibrated or measured form of punishment; and (3) of little or questionable accuracy in who and what it punishes" (Klonick, 2016, pp. 1029-1030).	This article discussed how internet communication changed the ways that social norm is enforced to regulate people's behaviour (i.e., via shaming). The ways to regulate online shaming by different social actors were also discussed.	Other	Online shaming is different from online bullying and harassment because of the element of norm enforcement. Nonetheless, as a punishment, it can still be problematic because it is indeterminate, inaccurate, and uncalibrated. To regulate online shaming, state regulation alone is not efficient or effective enough; Instead, Klonick suggested that legal, normative, and private remedies should be used along with state solutions.	Ethical/legal considerations, Other social actors, Retribution, Social/behavioural control

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
52	Laidlaw (2017)[a]	Law	Laidlaw (2017) suggested that shame can be understood as social control. And online shaming can be understood as a tool with a regulatory role. Laidlaw also proposed that online shaming involves different categories, that is, vigilantism, bullying, bigotry, and gossiping.	The research uses online shaming as an example to discuss the predominant debates around privacy.	Other	Through examining the dominant debates about privacy, Laidlaw differentiated online shaming from humbling, which is the instance "where an individual is rightly knocked down a peg for a social transgression" (p. 21). Online shaming, however, involves public humiliation that violates one's privacy (both public privacy and social privacy) and attacks their dignity. Laidlaw also identified a gap in privacy law. Although people's right to public privacy was protected under law in some countries, their right to social privacy, that is, the social dimension of privacy such as enabling participation in social spaces was not protected by law.	Abuse/ Stigmatisation, Ethical/legal considerations, Social/behavioural control
53	Larrain (2023)[a]	Feminist studies	Online shaming was described as a practice or strategy used by feminists for change.	This research examines *funa*, a specific type of shaming that addresses gender violence in Chile.	Semi-structured interviews (N = 32) with multiple different actors involved in *funa*	It was argued that *funa* as a feminist practice, offers "a problematic pathway to social change, which, despite contributing to denaturalising violence against women, does not address the structural causes of gender violence" (p. 80).	Addressing societal issue, Social justice, Retribution
54	Lauricella (2019)[a]	Education, Media and communication	Lauricella (2019) focused on a specific type of online shaming that targets students.	This article discussed the negative consequences of using student shaming in higher education.	Other	Student shaming can be very destructive to the student experiences and academic climate.	Consequences
55	Laywine (2021)[a]	Media and communication	Shaming as a digital activism by calling out those who do not engage in activism.	The author analysed a specific form of shaming, *Humanitarians of Tinder* (HoT), that shames people who post photos about their volunteer tourism on Tinder profiles.	Content analysis	It was argued that *Humanitarians of Tinder* (HoT) type of shaming demonstrates that the audience simultaneously engaged in justice-seeking as well as entertainment seeking, via establishing a moral community. However, the author questioned whether this practice can challenge the industries and the effectiveness of shaming for social change.	Addressing societal issue, Consequences, Effectiveness, Entertainment, Retribution, Social justice, Social belongingness, Value violation
56	Leopold et al. (2019)[a]	Management, Media and communication, Psychology	According to Leopold et al., online shaming enforces social norms and serves as a behavioural deterrent.	This paper discussed the effectiveness of the #MeToo social movement in changing the social norm.	Other	#MeToo social movement in which offenders of sexual harassment are publicly shamed has more effectively changed social norms than laws and organisational policies have done.	Addressing societal issue, Effectiveness, Ethical/legal considerations, Justifiability, Social/behavioural control
57	Loveluck (2019)[a]	Criminology	Loveluck defined digital vigilantism as "*direct online actions of targeted*	The goal of this article was to clarify the	Mixed-method of digital	Through discussing the typology of digital vigilantism, Loveluck suggested that digital	Consequences, Ethical/legal

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
			surveillance, dissuasion or punishment which tend to rely on public denunciation or an excess of unsolicited attention, and are carried out in the name of justice, order or safety" (p. 213) and "whereby individuals seem to be 'taking the law into their own hands' online'" (p. 214). Different sub-types of digital vigilantism were identified.	conceptual understanding of digital vigilantism and provide a typology for different types based on the different dimensions of digital vigilantism (i.e., trigger, target, motive, tactics, and organisational forms).	ethnography and content analysis	vigilantism is more than just personal revenge. It can convert "outrage, security concerns or assumptions of injustice into effective action online" (p. 24). However, as a powerful, informal punishment, it might also interfere with the legal system and cause potential violence.	considerations, Moral outrage, Retribution, Social justice, Social/behavioural control
58	MacPherson & Kerr (2020)[a]	Psychology	Public shaming was referred to as "practices by which an individual or a group communicates disapproval towards another individual in response to the person's transgression of a norm" (p. 1), for "the purpose of humiliation, social denouncement, and punishment" (p. 1).	This research aims to answer the following questions: When athletes are shamed by sport fans on social media for the athletes' legal, social, and sport-specific norm transgressions, is the shaming influenced by the athletes' gender? If so, what do these shaming practices look like?	Thematic/ Content analysis	Through examining 7700 comments made by sport fans on social media, the authors found that fans' online shaming practices in response to athletes' transgressions is gendered, as demonstrated in the content that includes objectification of females, victim blaming, and so on.	Value violation, Retribution
59	Mahmood et al. (2018)[a]	Law	Online shaming was defined as an instrument of social control. It was also classified as a form of "cyber bullying and cyber harassment" (p. 1127).	The study examined people's personal experiences of encountering shaming comments online.	Survey	Sexism, racism, and religious bigotry were found to be the most common behaviours that were associated with online shaming encountered by the participants online. It was found that 30.3% of the participants have liked, shared, or commented on a shaming post online.	Abuse/ Stigmatisation, Social/behavioural control
60	Mallén (2016)[a]	Criminology	Mallén (2016) suggested that online shaming can be understood as a kind of status degradation ceremony.	The research analyses a case of shaming that was triggered by posting a video clip. Online shaming exerted as a virtual punishment and eventually enabled a justice process that occurred online.	Thematic analysis	It was found that viewers of the film clip perceived it to be authentic and represent the truth of the incident, which enabled the process of shaming the customer who was portrayed as wronged in the film clip. However, Mallén suggested that the film clip only showed one of the alternative accounts of the incident.	Retribution, Social justice
61	Marwick (2021)[a]	Media and communication	Shaming was defined according to networked harassment. Specifically, morally motivated networked harassment (MMNH) was described as "a member of a social network or online community accuses an individual (less commonly a brand or organization) of	Through interviewing people who have experienced MMNH (n = 28) and workers at social media platforms (n = 9), it was analysed how moral outrage is	Semi-structured interview	The author proposed a model for morally motivated networked harassment (MMNH) model. The model involves identification of norm violation(s), justification of the harassment, and networked audience who promotes/amplifies on social media. Especially, the audience can share an	Consequences, Retribution, Social belongingness, Moral outrage, Justifiability, Value violation

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
			violating the networks' moral norms" (p 2.).	deployed to justify MMNH.		"ideological consensus of the accusing network", which is "reinforced through a common enemy and the symbolic boundaries between contexts are reinforced" (p. 5). Although the interviews showed that MMNH can be used by people with different political views, "people who challenge normative power structures (...) are more likely to be harassed by people who adhere to traditional social norms" (p. 2)	
62	Mielczarek (2018)[a]	Media and communication	Public shaming (campaign) is an aggregated punishment imposed on a norm violator. In the case study mentioned in this study, it can be done through the creation and circulation of internet memes.	This study examined the case of how people used internet memes to shame a police officer who appeared on the "pepper-spraying cop" image from the Occupy Wall Street movement at the University of California.	Iconographic tracking	People used internet memes as a weapon to seek social justice and engage in large-scale public shaming, that is, to punish and bully a transgressor.	Retribution, Social justice, Value violation
63	Milbrandt (2017)[a]	Sociology	Drought shaming is a type of civic online shaming that calls out water-wasters on social media. It was argued that this type of online shaming differs from naming and shaming (i.e., primary punitive) as most subjects of drought-shaming were not directly named or visibly identifiable and were not aware of themselves being shamed.	This article examined the case of drought-shaming in California (2014-2015), highlighting the role of images, which were taken non-consensually and circulated online.	Other	Four types of drought shaming were identified, based on the different shaming subjects (including public and private institutions, celebrities, and hyper-affluent property owners). According to Milbrandt, the drought was not only a natural disaster but also became a moral drama, where people used visual and discursive means to symbolise the water austerity as a moral, civic duty. And excessive water wasting became a representation of immorality.	Value violation, Justifiability
64	Moore (2016)[a]	Sociology	Naming and shaming as a punishment.	This article examined the media representation of four cases that were convicted of "gender fraud" in the UK.	Critical discourse analysis	It was found that the online news stories that reported "gender-fraud" cases and the accompanying readers' comments became the tool used in the shaming and humiliation punishments. These punishments were predicated on the gendered norms and signified what is (ab)normal within the society. The notions of what is socially (un)acceptable are also constructed and reinforced via these punishments.	Other social actors, Retribution, Social/behavioural control, Value violation
65	Muir et al. (2021)[a]	Media and communication	Online shaming can be described as "a phenomenon whereby individuals participate in social policing by shaming people on the internet over perceived violations of social norms or some other apparent wrongdoing" (p. 1).	This research examined how the phenomenon and concept of online shaming was framed in online news media.	Qualitative analysis	Through analysing the collected 69 news articles, it was found that online shaming was constructed as predominately negative and destructive, emphasising the severe consequences it can bring (such as abuse, ostracism, tragedies). Though there was a	Abuse/ stigmatisation, Consequences, Value violation

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
						smaller number of news reports highlighted how online shaming can lead to positive outcomes and be constructive (such as resulting people to rally behind the shamed, the shamed character reflected on the misdeed and learned from the mistake). These results therefore shows that many nuances and inconsistencies appear in shaming.	
66	Muir et al. (2023)[a]	Psychology	It was argued that there lacks a consensus on the definition of online shaming. However, it can be broadly understood as "social policing by shaming perceived transgressions via the internet" (p. 1).	This research examined how individual traits and characteristics (such as empathy, moral grandstanding, and psychopathy) can predict people's online shaming engagement.	Correlational online survey, Content analysis	It was found that the predictors (as shown the in the focus of this record) together accounted for 39% of variance in online shaming intentions, and 20% of variance explained in perceived deservedness of online shaming. An analysis of an open-ended question suggested that public perceived online shaming as involving "two sides to every story". For example, it can be a form of accountability but also have destructive effects. Others perceived it as a form an entertainment, hurting, as a social norm etc.	Consequences, Retribution, Social/behavioural control, Entertainment, Value violation
67	Murumaa-Mengel & Lott (2023)[a]	Media and communication, Sociology	Online shaming can be social sanctions that are reintegrative and disintegrative. Another form is recreational shaming, that is, "humour-based playful collective shaming that often occurs via online platforms" (p. 944).	This research analysed Facebook group that engaged in recreational shaming (n = 65) and the group organiser/ administrators (n = 8). This research investigated questions such as what they created, how they functioned, how the shaming enforced.	Content analysis, in-depth qualitative interview	Through the content analysis of Facebook recreational shaming groups and interviews with organisers, it was found that recreational shaming mainly motivated by "social belonging needs and entertainment gratification" (p. 944), in addition to functioning as reintegrative and disintegrative social sanctioning. These sanctioning practices can also be targeted at people who engage in shaming within the group (who violates the norms about shaming).	Entertainment, Social belongingness, Retribution, Social/behavioural control, Value violation
68	Norlock (2017)[a]	Philosophy	No explicit definition of online shaming was given.	Norlock discussed the crucial role of imaginal relationships with others involved in online shaming.	Other	Norlock suggested that engagement in social media increases our imaginal relationship with others, which entails more responsibilities of ours and a need for ethical assessment. Norlock suggested that it is important to consider the role of imagined relationships involved in online shaming (e.g., a shamer, i.e., a person who engages in shaming, might have the need for social recognition from fellow shamers). It was further argued that ethics recommendations grounded in these imaginal relationships can	Ethical/legal considerations, Social recognition/ approval

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
						help to reduce people's engagement in online shaming.	
69	Ong (2012)[a]	Law	Online vigilante justice or human flesh search was described as " the technology can be used to publicly shame, harass and humiliate a person with devastating effects" (p.127). It invovles the enforcement of social norms and values and can be seen as a way that people express dissatisfaction over justices in China.	This article examined the consequence of human flesh search (i.e., privacy infringement) and the legislation regarding this practice in China.	Other	This article examined a case where the Beijing Court recognised the practice of human flesh search used in China and recognised it as an infringement to the rights of privacy and reputation.	Consequences, Ethical/legal considerations, Social justice, Value violation
70	Oravec (2019)[a]	Political science	Online (social) shaming was described as a phenomenon that "involves the intentional collection and dissemination of data that are potentially stigmatizing in modes that are widely accessible and in which observers (including members of the public) can often add input" (p. 1).	This article discussed the emergence of online administrative shaming practices.	Other	In western countries (e.g., US, Australia), governmental and agency units are increasingly using shaming strategies to address social problems (e.g., recipients of public welfare, families that are behind in their school lunch payments). Those units focus on punitive exposure and stigmatizing information. However, such "insensitive ways of addressing social problems can foster" (p.17) fear, uncertainty, and the potential for reputational harm.	Addressing societal issue, Consequences, Ethical/legal considerations, Other social actors, Retribution
71	Packiarajah (n.d.)[b]	Criminology, Psychology	Packiarajah suggested that there was no conclusive definition of online shaming. It was suggested that the different forms of online shaming have not yet been comprehensively categorised in the literature. However, it was suggested that "online shaming is, at its heart, the perceived violation of a social norm by the offender" (p. 1). Therefore, norm enforcement was understood as one key aspect of online shaming.	This study examined the personal, environmental, and behavioural factors that might influence the victimisation and/or the perpetration of online shaming.	Online survey	Previous online shaming victimisation strongly predicted online shaming perpetration. However, other factors (i.e., age, gender, sexual orientation, internet self-efficacy, shame proneness, social comparison, social comparison status, perceived anonymity, social norm acceptance, masculinity, power distance, uncertainty avoidance, collectivism) were not significant predictors of online shaming victimisation or perpetration.	Consequences, Social/behavioural control, Value violation
72	Pan (2012)[b]	Sociology	Human flesh search was understood as a particular form of cyber surveillance that happens in China, "in which unrelated Internet users collaboratively conduct surveillance on fellow citizens" (p. 1). Although it was used as a practice for information gathering in China, in recent years, people have used it to identify, humiliate, shame and punish individuals.	This study examined why and how human flesh searches happened in China and explored the motives people have when they engage in such behaviour.	Content analysis, online survey, and in-depth interview	Pan found that the primary motive people reported was to help others, followed by having fun, Social justice, earning the virtual currency of the website, and making friends. Although some participants reported that they felt of being personally powerful, the majority reported a sense of collective empowerment as they believed the collective effort was powerful. It was also found that previous experiences in participating in the human flesh search engine increased participants' sense of empowerment.	Consequences, Effectiveness, Entertainment, Retribution, Social justice, Social recognition/ Approval

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
73	Papp et al. (2017)[a]	Psychology	The authors focused on a specific type of online shaming, that is, "slut-shaming". It was defined as a humiliating act that is based on a woman's presumed sexual behaviour and appearance.	This study examined how the target being "slut-shamed" and those who engaged in "slut-shaming" on Facebook were viewed by female college students. It also examined whether social class, clothing, and feminist identity affect people's acceptability of "slut-shaming" behaviour.	Survey	Although participants generally evaluated people who "slut-shamed" a target in a negative light, the target's attire had an effect on how women perceived people who engaged in shaming (i.e., shamer): participants were more willing to be closer to the shamer who shamed a provocatively dressed target than who shamed a conservatively dressed target. Attire and social class also affected how women perceive the "slut": It was also found that participants wanted more social distance from the provocatively dressed, high-SES target than the conservatively dressed, high-SES target. The shaming comment was interpreted as having a most serious tone when the target is from high SES and dressed provocatively. Lastly, feminist identity also found to play a role in influencing participants' perceptions about the "slut" and the shamer, with self-identified feminists being more willing to spend time with the "slut" and found the shaming act less justified than non-feminist participants did. (p. 240)	Abuse/ Stigmatisation, Justifiability, Value violation
74	Papp et al. (2015)[a]	Psychology	Slut-shaming, which can be understood as subtype of shaming.	This study examined whether people's acceptability and attitude of slut-shaming is affected by gender.	Online survey	Papp et al. (2015) found a sexual double standard that the gender of the shaming target engaging in "slut-shaming" affected how people judged the target. Male shaming targets were judged more harshly than females. Although people who engaged in shaming were generally evaluated negatively by the participants, they perceived the person who engaged in shaming as more judgemental and less admirable when the shaming target was a female rather than a male. Qualitative data indicated that people made different assumptions of the "sluts" based on the gender of the "sluts".	Abuse/ Stigmatisation, Justifiability
75	Pundak (2021)[a]	Psychology	Online public shaming campaigns, directed at individuals, brands, and firms, was perceived to have a potential to prevent future harm at a broader, societal level as well as causing harm for the wrongdoer who are shamed.	When and why do people engage in online shaming campaigns? Specifically, whether *Nonmaleficence* principle, (or the belief that one should avoid	Experiments	It was found that people who had a higher adherence to nonmaleficence principle were more likely to engage in online shaming, but only when the identifiability of the perceived wrongdoer was low. As when the identifiability is relatively low compared to high, the perceived harm inflicted on the	Consequences, Justifiability

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
				causing intential harm on others) might affect people's engagement in online shaming.		perceived wrongdoer due to shaming is expected to be milder.	
76	Puryear (2020)[b]	Psychology	Online shaming was not clearly defined in the thesis. However, Puryear (2020) suggested that people use the expression of outrage and shaming to condemn, punish moral transgressors in order to signal one's reputation and/or to coordinate others online for making changes (i.e., "to change minds, and to rally people to our side of a dispute", p. 2).	This research examined the possible motives and factors that influence people's outrage expression.	Online experiments and using the dictionary-based approach to analyse moral-emotional language	When participants were presented with viral, offensive comments, they felt stronger outrage and stronger desire to act than those who were presented with non-viral, offensive comments. The authors also found that virality strongly predicted people's expressed outrage and other negative moral emotions when encountering an opponent, but not when encountering people who hold a similar view. These results suggested that the primary drive for people to express their outrage online is to undermine the opponent's view rather than pursuing personal reputational rewards or social approval.	Moral outrage, Retribution, Social recognition/ approval
77	Puryear & Vandello (2019)[a]	Psychology	Puryear and Vandello described the hostility characterised by offensive language as flaming behaviour.	This research examined whether and when people would have more dull emotional responses to offensive speech when encountering online than offline.	Experiments	Compared to face-to-face interactions, when social information was lacking online (e.g., lacking profile picture and name of the victim), people felt less outrage towards the person who posted an inflammatory comment and had a lower intention to punish that person, as they were less surprised about the comment that was made online and perceived less harm towards the victim of the insulting comment.	Abuse/ Stigmatisation, Consequences, Justifiability, Moral outrage
78	Rost et al. (2016)[a]	Psychology, Sociology	Online firestorm was described as a type of online, collective aggression that targeted companies, public figures, and other individuals. However, Rost et al. suggested that it is different from other types of online aggression (such as cyberbullying and online harassment) and enforces social norms.	This research examined why people engage in an online firestorm and whether they engage in shaming out of the motive of enforcing social norms.	Mixed-method big-data approach	Rost et al. found preliminary supports for the proposed social norm theory that people engaged in a higher level of aggression when they were non-anonymous compared to being anonymous. Their engagement in online firestorms was also driven by intrinsic motivation (i.e., fairness concerns) and whether the instance of the online firestorm is in high controversy.	Abuse/ Stigmatisation, Justifiability, Soci al justice, Social recognition/ approval, Social/behavioural control
79	Sawaoka & Monin (2018)[a]	Psychology	Online shaming as a viral, moral outrage. It is also a form of aggregated punishment.	This research examined how people perceive and respond to those who engaged in viral outrage.	Online experiments	Sawaoka and Monin (2018) found that observers became more sympathised with the offender when the outrage became viral (i.e., aggregated punishments) compared to a non-viral condition. Sympathy was found to mediate the effect of virality on the observer's negative impression of the commenters. However, when the participants	Justifiability, Moral outrage, Retribution

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
						engaged in shaming themselves, they did not feel more sympathy when the outrage is viral than non-viral, nor did the feelings of sympathy affect their negative impression of the commenters.	
80	Sawaoka & Monin (2020)[a]	Psychology	Online shaming was understood as an expression of moral outrage triggered by the violation of moral standards. When the outrage becomes viral, it can be perceived as an excessive punishment.	This research examined how people perceive those who were the target of viral outrage.	Online experiments	It was found that when more people started to express outrage, those who observed it believed it was normal to do so. However, they also perceived the outrage as becoming excessive and felt sympathy toward the person who are shamed. These two processes (normative influence) and sympathy were found to suppress one another in predicting people's condemnation of the person who has violated a moral standard.	Justifiability, Moral outrage, Retribution, Value violation
81	Šincek (2021)[a]	Psychology	Online shaming was understood as a type of cyberviolence.	This research aimed to explore the psychometric properties of the revise scale of Committing and Experiencing Cyber-Violence Scale using an adolescent sample.	Online questionnaire	Exploratory factor analysis showed a five-factor model with satisfactory reliability, in which shaming was included as the first factor. Specifically, shaming included rude comments, gossiping, exclusion from groups etc.	Abuse/ Stigmatisation
82	Shenton (2020)[a]	Sociology	Based on the work of Herzfeld, online shaming was described as a process where identities are created through "which antagonists criticize one another" (p. 170), which can become polarised quickly.	This research examined the divisive nature of public online shaming via using memes, hashtags, and other posts.	Other	Shenton (2020) argued that online communities are formed via affirming the values that are endorsed by other insiders as well as constructing "otherness" that the insiders should oppose. Shaming materials are often constructed in a divisive way that the audience is encouraged to either agree or disagree. They are likely to circulate in like-minded individuals. Thus, the circulation of these shaming materials is likely to cause further division and polarisation.	Consequences, Social recognition/ approval, Value violation
83	Skoric et al. (2010)[a]	Criminology, Law, Psychology, Sociology	Skoric et al. (2010) described online shaming as a type of whistleblowing behaviour (prosocial behaviour) because they believe that those who engaged in online shaming were intending to report others who violated social norms. According to Skoric et al., people who engage in shaming are concerned with promoting civility and enforcing social norms.	This study examined why people engage in online shaming and how the individual differences (i.e., personality traits and endorsement of Asian values) influence the tendency of being deterred by online shaming and the tendency to contribute to shaming websites.	In-depth interview and survey	The interviews revealed that one reason for people to participate in shaming is to raise awareness about inconsiderable behaviour. However, people who participated in shaming were not purely altruistic (or prosocial). For example, it was found that personal negative experiences with bad behaviours affected their contribution to shaming. In the second study, it was found that people who were more likely to be deterred by online shaming, and people who were more likely to engage in online shaming showed differences in personality	Addressing societal issue, Value violation

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
						traits. Especially, people who were more likely to engage in online shaming seemed to be more socially responsible and scored higher on openness.	
84	Spring et al. (2018)[f]	Psychology	"Pile-on" (or viral online outrage) is a common response of experiencing outrage, that is, anger at the violation of moral standards that one holds.	The authors discussed the upsides of moral outrage. Specifically, it can be a critical force to collective action.	Other	Outrage is often described in a destructive way in the literature. The authors suggested that future research needs to study moral outrage in a way that bridges moral psychology and intergroup literatures. Both negative and positive consequences need to be studied.	Consequences, Moral outrage
85	Spring et al. (2019)[f]	Psychology	"Pile-on" (or viral online outrage) is a common response of experiencing outrage, that is, anger at the violation of moral standards that one holds.	This article is a response to Brady and Crockett (2019).	Other	The authors argued that the experience of outrage is different from the expression of outrage. Although expression of online outrage might cause negative consequences, it does not mean experiencing outrage would have more downsides than upsides.	Consequences, Moral outrage, Value violation
86	Suhaimi et al. (2018)[a]	Law	The authors suggested that based on the functions of online shaming, it can be described as both an internet vigilantism and a social/behavioural control in society via reward and punishment.	This study examined people's perceived efficacy of online shaming as a form of social control.	Survey	It was found that the majority of the participants believed that the efficacy of online shaming as social control is impaired by online abuse and thus undermines the social order.	Abuse/ Stigmatisation, Effectiveness, Ethical/legal considerations, Retribution, Social/behavioural control
87	Sundén & Paasonen (2018)[a]	Feminist studies, Media and communication	Shaming was described as a tactic used in online hate to silence others.	This article explored "shamelessness" as a tactic used by feminists in a Nordic context for resisting online hate and public shaming.	Other	"Shamelessness" is used to fight back misogynist and online hate through reinterpreting the shaming discourse, though the outcome is uncertain as it is based on the volatility of affects.	Abuse/ Stigmatisation, Consequences
88	Tandoc et al. (2022)[a]	Media and communication	The in-depth interviews provided insights on how users on social media understand cancel culture and its motivations. Cancel culture was defined as a practice that involves a group of people shaming a target publicly on social media, with a potential aim to "hold the target accountable for socially incorrect or unacceptable behavior" (p. 9). Different motivations were identified, including educating others on certain social issues, seeking accountability, seeking a sense of justice through punishment, correcting the power imbalance etc.	The research involved conducting in-depth interviews to examine how cancel culture is understood. Survey was conducted to examine the predictors of people's intention to cancel others, based on the Theory of Planned Behaviour.	In-depth interviews, survey	See the definition part for the results of the in-depth interviews. Aligning with the Theory of Planned Behaviour, it was found from the survey that people's attitude, subjective norms (including both descriptive and injunctive norms) and perceived behavioural control positively predicted people's intention to engage in cancel culture. Specifically, perceived behavioural control was found to be the strongest predictor among others, and injunctive norm was a stronger predictor than descriptive norm, suggesting a sense of obligation. While general belief in a just world negatively predicted people's intention, and	Retribution, Social justice, Social/behavioural control

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
						personal belief in a just world was not a significant predictor. Therefore, cancel culture seems to be closely related to desire for social justice.	
89	Thompson & Cover (2022)[a]	Media and communication	This research understood online shaming and internet pile-on as one form of online hostility.	Through analysing a number of Twitter posts (N = 60) on a specific case of online shaming that led to the suicide of the victim, this study examined what is involved in the online hostility.	Case study	Overall, the discourses involved in the online hostility emphasised the shortcomings of the victim, rather than how problematic that the online hostility was, suggesting that the public discussion ignored the significance of hostility of online shaming and internet pile-ons.	Abuse/ Stigmatisation, Consequences
90	Trottier (2018)[a]	Criminology, Media and communication	Online shaming was defined as a form of mediated shaming where users scrutinise and publicly expose others for punitive denunciations.	This research discussed the relationship among different social actors who perform shaming.	Other	It was argued that social actors (i.e., press and state) played an important role in the specific form of user-led surveillance, online shaming, by stigmatising and excluding "(categories of) individuals under scrutiny" (p. 170). Specifically, shaming can be used by social actors to mobilise others. In other words, it can be understood as a social control, which is subjective to the reinforcement of "discrimination and categorical struggles over legitimate use of public space" (p. 178).	Retribution, Social/behavioural control, Other social actors
91	Trottier (2020a)[a]	Media and communication	Shaming was described as a form of digital vigilantism. Digital vigilantism was defined as practices that "scrutinise, denounce and even leverage harm against those deemed to transgress legal and/or moral boundaries, with the intention of achieving some form of justice." (p. 197) It was suggested that media could play a crucial role in facilitating digital vigilantism, and that this phenomenon should be understood accordingly.	This paper examines how digital vigilantism is made meaningful. Particularly, the author put forward a conceptual model that emphasises the "coordinated, moral and communicative components" (p. 196) of digital vigilantism.	Other	It was argued that (the relations between) various social actors (i.e., campaign initiators and the shaming target, participants and the states, the press, and social media) can facilitate or even contribute to digital vigilantism. Despite the intention to achieve justice, digital vigilantism could reproduce inequalities and/or discriminations.	Value violation, Retribution, Social justice, Other social actors
92	Trottier (2020b)[a]	Media and communication	Digital vigilantism was defined as a type of citizen-led justice-seeking. Mediated denunciation and shaming are practices used in digital vigilantism.	This paper explored how digital vigilantism (e.g., denunciation) was expressed by the UK press as well as the coverage on the motivations of why people participate or facilitate denunciations,	Case study	It was found that the UK press might have played a crucial role in online denunciations. Although news coverage on digital vigilantism was seemingly neutral, it was imbued with the understandings of the agent of news (e.g., journalists, the stance of the press), while lacking an acknowledgement of its role as a denouncer. Regarding the Press's coverage on the motivations to name	Value violation, Retribution, Social justice, Other social actors

	Author (Year)	Discipline	Definition of online shaming	Focus	Method	Key findings/arguments	Initial classification
				through analysing five incidents of digital vigilantism.		and shame, many reasons can be identified, including the offences that are "generalised to broader causes or concerns" (p. 607), the socio-political context that facilitated the denunciations, vulnerabilities of those who are targeted and/or the denunciators, as well as a mobilisation based on the value that triggered the denunciation in the first place. Therefore, the press played a crucial role in making digitally mediated shaming (i.e., online shaming) socially meaningful.	
93	Wall & William (2007)[a]	Criminology	Shaming was understood as a tactic employed by vigilante groups. It is one method of online governance.	This article explored and examined the ways that online 'communities' maintain order.	Other	Drawing on Braithwaite's theory, the authors distinguished reintegrative and disintegrative shaming tactics of online vigilantism. According to the authors, the effectiveness of reintegrative shaming to deter offenders is complicated by the social bonds of the offender with other members within the online community, as it is unlikely to be effective when the offender lacks online interdependencies with other members. Although disintegrative shaming can ostracise the offender from the community permanently, the effectiveness is still uncertain because of the anonymous and temporary nature of such punishment. Rather, Wall and William suggested that it could be worthy to combine online vigilantism (i.e., online shaming and humiliation) with organised policing.	Effectiveness, Retribution, Social/behavioural control
94	Wehmhoener (2010)[b]	Media and communication	Cyber/public shaming was described as a type of internet vigilantism.	This thesis aimed to understand the phenomenon of internet vigilantism and people's attitude towards it through analysing a specific case study.	Thematic analysis	It was found that most people wanted to punish the transgressors and achieve justice. Four themes were identified in how people called for action: 1) most people made moral condemnation of the transgressors 2) the call to action was requested directly by people 3) most calls to action were on punishment and 4) people use calls to act collectively to make requests.	Retribution, Social justice

Categorisation Process for Examining Researcher's View on Online Shaming

The included articles were divided into 5 different categories based on the initial classifications accompanying the articles. These categories are: 1 = Punishing the perceived wrongdoer (n = 24), 2 = Deterring the perceived wrongdoing with punishment (n = 19), 3 = Seeking social acknowledgement with punishment (n = 6), 4 = Creating change with punishment (n = 7), and 5 = Seeking deterrence, social acknowledgement and/or change (n = 20). Specifically, the categories formed as:

Category 1 includes articles with initial classifications with "Retribution" or "Abuse/Stigmatisation", but excludes "Social/behavioural control", "Social approval/recognition", "Addressing societal issue", "Social belongingness", and "Social justice". We suggest that this category captures articles that emphasise online shaming punishment can reflect motives underlying the goal of punishing the perceived wrongdoer.

Category 2 includes articles with initial classification "Retribution" or "Abuse/ Stigmatisation", and "Social/behavioural control" or "Social justice", but excludes "Social approval/recognition", "Addressing societal issue", and "Social belongingness". We suggest that this category captures articles that emphasise online shaming punishment can reflect motives underlying the goal of deterring the perceived wrongdoer.

Category 3 includes articles with initial classification "Retribution" or "Abuse/ Stigmatisation", and "Social belongingness" or "Social approval/recognition", but excludes "Addressing societal issue". We suggest that this category captures articles that emphasise online shaming punishment can reflect motives underlying the goal of seeking social acknowledgement.

Category 4 includes articles with initial classification "Retribution" or "Abuse/ Stigmatisation", and "Addressing societal issue". We suggest that this category captures

articles that emphasise online shaming punishment can reflect motives underlying the goal of creating change.

Category 5 includes articles that includes “Social belongingness”, “Social approval/recognition”, “Social justice”, or “Addressing societal issue”, but excludes “Punishment” and “Abuse/Stigmatisation”. We suggest that this category captures articles that emphasise online shaming as driven by deterrence for justice, acknowledgement, and/or social change without understanding it as a punishment.

Appendix B: Study 2 Supplementary Materials

Data Collection

Our data collection spanned from Sunday 8 March 2020 to Thursday 12 March 2020, as the event was still evolving. To gather as many tweets as possible, we combined two different methods of collecting tweets using the *rtweet* R package (Kearney et al., 2022) and Twitter's API. The first method was the *search_tweets* function, which allows us to collect tweets matching our search query and has been posted in the past 6-9 days. On 8 March 2020, we used the following keywords[11] for the archival tweets: *Melbourne gp*, *Toorak gp*, *melbourne doctor*, *Toorak doctor*, *dr Higgins*, *flabbergasted*, *#istandwithChrisHiggins*, *flabbergaslighting*, *Jenny Mikakos*, and *@JennyMikakos*. The initial search acquired 30,783 unique tweets. We also used *stream_tweets* to acquire a live stream of tweets. The tweets were streamed using the following keywords: *Melbourne gp*, *Toorak gp*, *melbourne doctor*, *Toorak doctor*, *dr Higgins*, *#flabbergasted, #IStandwithChrisHiggins*, *flabbergaslighting*, *Jenny Mikakos*, *@JennyMikakos*, *#IStandWithJenny*[12]. The initial search acquired 3,021 unique tweets from two attempts of streaming tweets that together spanned 48 hours.

Data Cleaning

Tweets that were created before 2020-03-07 00:10:27 UTC time (or 2020-03-07 11:10:27 AEDT time, as the approximate time that the news regarding the GP infecting

---

[11] According to the Twitter API documentation, the queries were not case-sensitive (i.e., *Melbourne gp* will match *Melbourne gp*, *Melbourne Gp*, *Melbourne GP*, *melbourne gp*, *melbourne Gp*, *melbourne GP*): https://developer.twitter.com/en/docs/twitter-api/tweets/search/integrate/build-a-query#punctuation

[12] The queries we used for streaming tweets differed from the queries we used for searching tweets, as the primary researcher browsed the tweets collected from the searched tweets and updated the search queries for streamed tweets. For example, some people used *#IStandWithJenny as a response to #IStandWithChrisHiggins*, hence the hashtag has been added to the search queries, whereas *flabbergasted* introduced a large portion of irrelevant tweets (i.e., noise), hence the keyword has been replaced with *#flabbergasted*.

patients) were excluded, leaving us 11,733 searched tweets. Irrelevant tweets were removed[13], leaving us 10,389 searched tweets and 3,021 streamed tweets. After removing retweets, 3,496 searched tweets and 1,639 streamed tweets remained. We further excluded tweets that were posted by media outlets, as we are interested in why individuals engage in online shaming (but less on the role of media outlets). A total of 5,005 tweets (including 3,369 searched tweets and 1,636 streamed tweets) created by 2,248 unique users comprised the final corpus. Data was pre-processed by removing URLs, usernames, and emojis. Before removing stop words, numbers, and punctuations, we first spelt out the common abbreviations that were used frequently in the corpus. For example, "U.S." was replaced with "USA", and "gov" or "govt." was replaced with "government".

[13] To screen out the irrelevant tweets that *flabbergasted* introduced, tweets that mentioned *flabbergasted* will be included only if at least one the following keywords were also mentioned: *gp*, *jenny*, *health*, *covid*, *coronavirus*, *doctor*, *melbourne*, *vic*, *higgins*, *toorak*, *mikakos*, *victoria*, *#flabbergasted*. A total of 1,344 searched tweets were identified as irrelevant, whereas none of the streamed tweets was identified as irrelevant.

Appendix C: Study 3 Supplementary Materials

Pilot Study Results

Table C1

Mean Scores of Shaming Scenarios

Scenarios	Hostility ($N = 27$)	Hate speech ($N = 33$)
1	3.48(1.12)	2.70(1.53)
2 (Halloween)	3.30(1.17)[a]	2.94(1.46)
3 (Homelessness)	3.41(0.89)[a]	2.82(1.29)
4	3.59(1.12)	2.70(1.36)
5	4.00(0.88)	2.94(1.39)
6 (Dog in snow)	3.37(0.84)[a]	2.67(1.24)
7	3.00(0.96)	2.55(1.35)
8 (Bear in zoo)	3.41(1.12)[a]	2.61(1.32)

Note. Standard deviations are presented in parentheses.
[a] Scenarios 2, 3, 6, 8 were used in Study 3.

Table C2

Mean Scores of Shaming Comments

No.	Comments	*N*	Mean (*SD*)
1	What you said is wrong.	30	1.53(0.78)
2	This behaviour is unacceptable!	32	1.69(0.82)
3	Really? You think this is a right way to respond?	31	1.81(0.83)
4	That was uncalled for	30	1.43(0.57)
5	I'm disturbed by your behaviour.	32	2.00(0.84)
6	You have taken this too far!	30	1.83(0.91)
7	Stop saying things like that! It's toxic	32	2.09(0.89)
8	Shame on you for doing that. You should know better.	32	2.25(0.88)
9	can't believe someone would do something as horrible as this	32	2.34(1.04)
10	Ahh what you did is upsetting!	29	1.69(0.89)
11	Your behaviour shows an underdeveloped level of maturity.	31	2.61(1.02)
12	I'm horrified by your terrible, terrible act	28	2.61(0.83)
13	I'm disgusted at your behaviour.	32	2.69(0.97)
14	Your behaviour makes me sick!	31	2.84(1.04)
15	Stop being abusive like this! It's bullying	31	2.00(1.03)
16	People like you are so uneducated.	29	3.00(1.04)
17	This person cannot go unpunished	31	2.65(0.99)
18	You should be ashamed of yourself.	30	2.73(1.02)
19	You are evil.	31	3.35(1.08)
20	You make me sick	28	2.93(0.86)
21	What the hell is wrong with you?	30	3.20(1.00)
22	This only proves what a sick person you are.	32	3.28(0.96)
23	What an idiot, honestly.	29	3.07(1.07)
24	Ignorant people like you are what's wrong with this world!	31	3.52(1.03)

25	Let's find out who this sick person is and punish him!	29	3.69(1.07)
26	You are a piece of trash!	32	3.88(0.94)
27	Pure TRASH...	25	3.68(1.03)
28	A Real PIG	31	3.32(1.05)
29	You're such a nasty person and deserve to be punished!!	30	3.47(0.97)
30	Send this idiot to prison.	28	3.71(1.12)
31	You're f*cking disgusting.	29	4.52(0.79)
32	Get a job! You're gross!	31	3.39(0.84)
33	Your face sickens me!	29	3.86(0.83)
34	Such quality human being #scum	30	3.47(0.97)
35	What an ignorant loser!	32	3.47(0.98)
36	You're a disgusting fool! You should've known better.	26	3.62(0.98)
37	Inexcusable. You're such a selfish, disrespectful brat!	30	3.70(0.88)
38	You are the utter disgrace.	32	3.44(1.05)
39	You're a scum and I'm offended by your existence	29	4.10(0.72)
40	You're a danger and should be put in a cage	31	3.71(1.10)
41	You are a threat to humanity	27	3.67(0.83)
42	You're no better than a skunk!	26	3.19(0.98)
43	You deserve to be banned permanently in our society.	26	3.81(0.90)

Note. Standard deviations are presented in parentheses. A total of 33 participants rated the shaming comments, yet all 43 shaming comments have missing values.

Study 3 Material
Mobilising identity condition:

National Institute of Online Hostile Speech Research and Prevention

Message from CEO

Thank you for showing your interest in this Artificial Intelligence (AI) training program.

I am **Dr. Taylor Jones, the CEO of a Non-Profit Organization** called National Institute of Online Hostile Speech Research and Prevention. As a charity organization based in the United States, **our goal is to build a hostility-free online environment with the aid of advanced technology.**

Anyone can become the target of online hostility (online bullying; trolling). In the previous incidents, the targets of online hostility were at the risk of self-harm, substance abuse, emotional abuse and even suicide. It is very important for every one of us to develop systems that can reduce online hostility so that we can all interact safely online together. This current project is testing whether we develop AI systems that are able to stop people from online hostility.

We need your help to train this AI system to detect and deter online hostility.
Our research has suggested a new theory that responding with disapproving comments towards those engaged in hostility online can be an effective way of educating them and stopping the hostile activity, such that we know how the program SHOULD respond to people who behave hostile online. You are important in helping us to train an effective AI system that implements the new theory.

We therefore ask you to teach the AI how to respond to online hostile comments. In this way, you are helping to keep people safe online.

For your information, the AI system will be first launched in the United States after it is well-developed. Please help us —— together we can make the online environment a better place!

Dr. Taylor Jones
CEO, National Institute of Online Hostile Speech Research and Prevention

Non-mobilising condition:

Message from CEO

Thank you for showing your interest in this Artificial Intelligence (AI) training program.

I am Dr. Wei Zhang, the CEO of a private company called China AdTech. As an IT company based in China, our goal is to become one of the best global advanced technology companies with top-selling AI products.

Online hostility can cause annoyance. Therefore, an AI system that can help to address online hostility has the potential to become a popular product, which is very important for us to become the best technology company. This current project is testing whether we develop AI systems that are able to address online hostility.

We need your help to train this AI system to address online hostility. You will be asked to teach the AI how to respond to online hostile comments. You are important in helping us to develop a top-selling AI product and become one of the best technology companies.

For your information, the AI system will be first launched in China after it is well-developed. Please help us —— together we can make China AdTech Inc. the best technology company!

Dr. Wei Zhang
CEO, China AdTech Inc.

Trials Presented in Study 3
First trial in the normal order/Last trial in the reversed order

Please choose one from the following comments to respond to the person who commented under the post:

What you said is wrong.
This comment is unacceptable!
Ahh what you said is upsetting!
AI should not respond

Second trial in the normal order/Third trial in the reverse order

Please choose one from the following comments to respond to the person who commented under the post:

Your comment shows an underdeveloped level of maturity.
I'm horrified by your terrible, terrible comment
This person cannot go unpunished
AI should not respond

Third trial in the normal order/Second trial in the reversed order

Please choose one from the following comments to respond to the person who commented under the post:

You are the utter disgrace.
You're such a nasty person and deserve to be punished!!
Ignorant people like you are what's wrong with this world!
AI should not respond

Last trial in the normal order/First trial in the reversed order

Please choose one from the following comments to respond to the person who commented under the post:

You are really a piece of trash! #scum
You're f*cking disgusting.
You’re a scum and I’m offended by your existence
AI should not respond

Study 3 Additional Measures

We measured the following variables but did not provide the details of them in the chapter: Participants were asked to respond to the questions about the identification with the commenter, perceived deservingness of the commenter, and emotions towards the commenter (i.e., outrage, gloating, and satisfaction), their moral position and conviction about preventing online hostility, their general view about AI, if they felt compelled to respond in the AI training, and the purpose of the study.

Perceived Responsibility. Two items were used to measure perceived responsibility to prevent online hostility: "It is my responsibility to prevent online hostility." and "I feel obliged to prevent online hostility.", $r = .86$, $p < .001$. Responses were given on a 7-point scale anchored 1 = *Strongly disagree* to 7 = *Strongly agree*. Participants' scores were averaged to create a single scale, with higher scores reflecting higher perceived responsibility to prevent online hostility.

Sense of Contribution. Two items were used to measure sense of contribution: "I feel that I have made a contribution to the issue of preventing online hostility." and "I feel bond to those who made a contribution to the issue of preventing online hostility.", $r = .64$, $p < .001$. Responses were given on a 7-point scale anchored 1 = *Strongly disagree* to 7 = *Strongly agree*.

Task Comfort. Two items, adapted from Haslam et al. (2014), were used to measure comfortability of continuing the AI training: "I found it hard to continue to the end of this AI training task." [Recoded] and "I was comfortable continuing to the end of this AI training task.", $r = .73$, $p < .001$. Responses were given on a 7-point scale anchored 1 = *Strongly disagree* to 7 = *Strongly agree*. Participants' scores were averaged to create a single scale, with higher scores reflecting more comfortable continuing to the end of the AI training.

Identification with the Commenter. One item, adapted from Postmes, Haslam, and Jans (2013) was used to measure identification with the commenter in the last trial of AI training, "I identify with the person who commented under the original post in the last trial.", on a 7-point scale anchored 1 = *Strongly disagree* to 7 = *Strongly agree*.

Perceived Deservingness. Three items, adapted from Woodyatt et al. (2017), were used to measure the perceived deservingness of the commenter to be shamed who commented in the last trial of AI training. An example of the perceived deservingness item is "The person who commented under the original post deserves to be called out for their behavior in a public way.", $\alpha = .72$. Responses were given on a 7-point scale anchored 1 = *Strongly disagree* to 7 = *Strongly agree*. Participants' scores were averaged to create a single scale, with higher scores reflecting higher perceived deservingness of the commenter in the last trial of AI training.

Emotions. Each of the emotions (outrage, gloating, and satisfaction) comprises 2 items. Example of an outrage item is "I feel outraged about the action of the person who commented under the original post.", $r = .82$, $p < .001$. Example of a gloating item is "I felt amused by responding to the person who commented under the original post.", $r = .78$, $p < .001$. An example of satisfaction item is "I felt satisfied by responding to the person who commented under the original post.", $r = .90$, $p < .001$. Responses were given on a 7-point scale anchored 1 = *Strongly disagree* to 7 = *Strongly agree*. Participants' scores on each of the emotions were averaged to create a single scale, with higher scores reflecting stronger emotions felt towards the commenter in the last trial the AI training.

Moral Position. One item was used to measure moral position on online hostility, "To what extent are you against online hostility?", on a 5-point scale anchored 1 = *None at all* to 5 = *Extremely*, with higher scores reflecting stronger moral position against online hostility.

Moral Conviction. Four items, adapted from Reifen Tagar, Morgan, Halperin, and Skitka (2014) and Skitka, Hanson, Washburn, and Mueller (2018), were used to measure moral conviction on online hostility. An example of the moral conviction item is "To what extent is your position on online hostility a reflection of your core moral beliefs and convictions?", $\alpha = .91$. Responses were given on a 5-point scale anchored 1 = *None at all* to 5 = *Extremely*. Participants' scores on the items were averaged to create a single scale, with higher scores reflecting stronger moral conviction on online hostility.

General View About AI. Two items, adapted from Zhang and Dafoe (2019), were used to measure the general view about AI: "I support the development of AI." and "I think the impact of AI has been more good than bad.", $r = .75$, $p < .001$. Responses were given on a 7-point scale anchored 1 = *Strongly disagree* to 7 = *Strongly agree*. Participants' scores on the items were averaged to create a single scale, with higher scores reflecting more positive view about AI.

Feeling of Being Compelled. Two items were used to measure how compelled people felt to respond in the AI training: "I felt compelled to choose from the comments because I am paid for this AI training program." And "I felt compelled to contribute my own comment because I am paid for this AI training program.", $r = .54$, $p < .001$. Responses were given on a 7-point scale anchored 1 = *Strongly disagree* to 7 = *Strongly agree*.

Social Media Platforms. One item was used to measure the social media platforms participants use: "Please select the social media platforms you use (you may select more than one)" (response options: Facebook, Twitter, Instagram, Tumblr, Reddit, YouTube, Online Newspapers, Other). Facebook (87.5%) and YouTube (80%) were two the most popular

social platforms, followed by Instagram (59.5%), Twitter (56.5%), Reddit (52.0%), Online Newspapers (23.5%), Tumblr (9.5%), and other[14] (5%).

Comprehension Checks. To ensure the participants paid attention and comprehended our manipulations, they were asked to answer four questions after reading the message from CEO. Three of them were multiple-choice questions: "Where does the CEO come from?" (response options: Australia, Korea, China, US), "Who organized this training program?" (response options: a government department, a university, a non-profit organization, an IT company), and "What is the goal of the organizers behind this training program?" (response options: using the training program to recruit specialists for 5G technology development, developing a top-selling AI product that helps the company to become the best in the world, building a safer online environment that is hostility-free, educating children about the importance of studying a programming language). Most participants answered the multiple-choice apprehension checks correctly, while 30.5% of the participants answered one of them incorrectly, 9.5% answered two of them incorrectly, and 2.0% answered all of them incorrectly.

[14] Other social media platforms include forums (1%), Goodreads (0.5%), Pinterest (0.5%), Reddit (0.5%), Snapchat (0.5%), TikTok (0.5%), Tinder (0.5%), Twitch (0.5%), WhatsApp (0.5%).

Table C3

Logistic coefficients for predicting shaming choice by trial

Predictor	*B* (*SE*)	Wald[a]	Odds Ratio	OR 95% CI
Outcome Variable: Trial 1 shaming choice[b]				
Constant	0.21 (0.29)	0.72	1.23	[0.70, 2.16]
Order[c]	1.21 (0.33)	3.67***	3.34	[1.76, 6.37]
Condition[d]	-0.88 (0.33)	-2.64**	0.42	[0.22, 0.80]
χ^2 (2, N = 174) = 21.95, p < .001, Pseudo R^2 = 0.12.				
Outcome Variable: Trial 2 shaming choice[b]				
Constant	1.41 (0.33)	4.27***	4.09	[2.14, 7.81]
Order[c]	-0.76 (0.34)	-2.24*	0.47	[0.24, 0.91]
Condition[d]	-0.36 (0.34)	-1.07	0.70	[0.36, 1.35]
χ^2 (2, N = 174) = 6.01, p < .05, Pseudo R^2 = 0.03.				
Outcome Variable: Trial 3 shaming choice[b]				
Constant	-0.63 (0.29)	-2.16*	0.53	[0.30, 0.94]
Order[c]	0.02 (0.32)	0.05	1.02	[0.54, 1.90]
Condition[d]	0.01 (0.32)	0.02	1.01	[0.54, 1.88]
χ^2 (2, N = 174) = 0.003, p = .999, Pseudo R^2 < 0.001.				
Outcome Variable: Trial 4 shaming choice[b]				
Constant	-1.33 (0.37)	-3.60***	0.26	[0.13, 0.54]
Order[c]	-1.16 (0.51)	-2.28*	0.31	[0.12, 0.85]
Condition[d]	-0.26 (0.47)	-0.55	0.77	[0.31, 1.93]
χ^2 (2, N = 174) = 5.95, p = 0.051, Pseudo R^2 = 0.03.				

Note. OR stands for odds ratio. [a] All Wald Z-tests presented in this table has a degree of freedom of 1. [b] Shaming choice by trial was coded as 0 = Choose to not respond, 1 = Choose to respond. [c] Order was coded as 1 = Normal, 2 = Reversed. [d] Condition was coded as 1 = Ingroup, 2 = Outgroup.
* $p < .05$. ** $p < .01$.

Table C4

Multinomial logistic coefficients for predicting trial 1 contributed comments

Predictor	Trial 1		
	B (*SE*)	Wald	OR [95% CI]
	Disagreeing vs. Non-punitive/No responses		
Condition[a]	-0.96 (0.57)	2.86	0.38 [0.13, 1.17]
Order[b]	-0.92 (0.60)	2.34	0.40 [0.12, 1.29]
Condition*Order	0.49 (0.86)	0.31	1.63 [0.30, 8.83]
Constant	< 0.01 (0.43)	< 0.01	NA
	Online shaming vs. Non-punitive/No responses		
Condition[a]	-1.00 (0.52)	3.69	0.37 [0.13, 1.02]
Order[b]	-0.60 (0.52)	1.30	0.55 [0.20, 1.53]
Condition*Order	1.29 (0.70)	3.42	3.64 [0.92, 14.30]
Constant	0.31 (0.40)	0.61	NA

Note. $N = 174$. OR stands for odds ratio. A test of the model with all predictors against a constant-only model was not significant, χ^2 (6) = 8.91, $p = .18$.
[a] Condition was coded as 1 = Ingroup (the reference group; n = 80), 2 = Outgroup (n = 94). [b] Order was coded as 1 = Normal (the reference group; n = 86), 2 = Reversed (n = 88).

Table C5

Multinomial logistic coefficients for predicting trial 2 contributed comments

Predictor	Trial 2		
	B (*SE*)	Wald	OR [95% CI]
	Disagreeing vs. Non-punitive/No responses		
Condition[a]	-0.16 (0.81)	0.04	0.85 [0.17, 4.18]
Order[b]	0.52 (0.77)	0.46	1.68 [0.37, 7.55]
Condition*Order	0.36 (1.03)	0.12	1.43 [0.19, 10.7]
Constant	-1.95 (0.62)	9.92**	NA
	Online shaming vs. Non-punitive/No responses		
Condition[a]	-0.53 (0.49)	1.19	0.59 [0.23, 1.53]
Order[b]	-0.25 (0.50)	0.26	0.78 [0.29, 2.06]
Condition*Order	0.73 (0.69)	1.12	2.07 [0.54, 7.98]
Constant	-0.48 (0.35)	1.85	NA

Note. $N = 174$. OR stands for odds ratio. A test of the model with all predictors against a constant-only model was not significant, χ^2 (6) = 3.47, $p = .75$.
[a] Condition was coded as 1 = Ingroup (the reference group; n = 80), 2 = Outgroup (n = 94). [b] Order was coded as 1 = Normal (the reference group; n = 86), 2 = Reversed (n = 88).
** $p < .01$

Table C6

Multinomial logistic coefficients for predicting trial 4 contributed comments

Predictor	Trial 4		
	B (*SE*)	Wald	OR [95% CI]
	Disagreeing vs. Non-punitive/No responses		
Condition[a]	-0.59 (0.57)	1.06	0.56 [0.18, 1.70]
Order[b]	-1.50 (0.69)	4.75*	0.22* [0.06, 0.86]
Condition*Order	1.50 (0.87)	2.99	4.50 [0.82, 24.7]
Constant	-0.37 (0.43)	0.72	NA
	Online shaming vs. Non-punitive/No responses		
Condition[a]	-0.84 (0.51)	2.69	0.43 [0.16, 1.17]
Order[b]	-0.84 (0.51)	2.69	0.43 [0.16, 1.17]
Condition*Order	0.47 (0.72)	0.42	1.60 [0.39, 6.60]
Constant	0.14 (0.38)	0.14	NA

Note. $N = 174$. OR stands for odds ratio. A test of the model with all predictors against a constant-only model was not significant, χ^2 (6) = 9.60, $p = .14$.
[a] Condition was coded as 1 = Ingroup (the reference group; n = 80), 2 = Outgroup (n = 94). [b] Order was coded as 1 = Normal (the reference group; n = 86), 2 = Reversed (n = 88).
* $p < .05$. ** $p < .01$

Appendix D: Study 4 Supplementary Materials

Ingroup Leader, Leader's Norm and Goal Presented:

Message from CEO

Thank you for showing your interest in helping us evaluate the materials used for a public health campaign about coronavirus.

I am Dr. Taylor Jones, the CEO of a Non-Profit Organization called the National Institute of Public Health Research. We are a charity organization based in the United States. We work on a public health campaign regarding COVID-19.

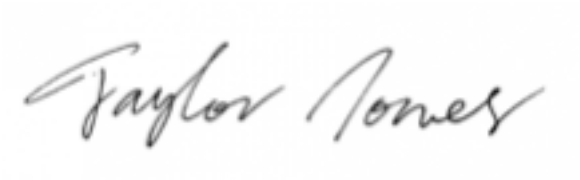
To protect yourself and others, people need to adhere to the guidelines of keeping a social distance from others, practicing good hygiene such as washing hands correctly and staying at home while sick, as well as following the limits for public gatherings.

The purpose of our campaign is to encourage people to adhere to the guidelines. All people should follow these guidelines, as they will not only flatten the coronavirus curve and protect our community, but will also help save the lives of those who are more vulnerable.

However, some people are not complying with the guidelines. This current survey is testing some materials that will be used for a public health campaign.

We need your help to launch the campaign. Our research has suggested a new theory that responding with disapproving comments towards those who did not follow the guidelines can be an effective way of educating people about how to behave appropriately. We ought to protect our community and educating non-followers may reinforce them to follow the guidelines. Therefore, you are important in helping us to evaluate whether the disapproving comments can teach non-followers to behave more carefully. In this way, you are helping to keep people safe.

For your information, the public health campaign combating COVID-19 will be launched in the United States after it is evaluated. Please help us —— together we can better protect the community at large!



Dr. Taylor Jones
CEO, National Institute of Public Health Research (US)

Outgroup Leader, Leader's Norm and Goal Presented:

Message from CEO

Thank you for showing your interest in helping us evaluate the materials used for a public health campaign about coronavirus.

I am Dr. Wei Zhang, the CEO of a Non-Profit Organization called the Chinese Institute of Public Health Research. We are a charity organization based in China. We work on a public health campaign regarding COVID-19.

To protect yourself and others, people need to adhere to the guidelines of keeping a social distance from others, practicing good hygiene such as washing hands correctly and staying at home while sick, as well as following the limits for public gatherings.

The purpose of our campaign is to encourage people to adhere to the guidelines. All people should follow these guidelines, as they will not only flatten the coronavirus curve and protect our community, but will also help save the lives of those who are more vulnerable.

However, some people are not complying with the guidelines. This current survey is testing some materials that will be used for a public health campaign.

We need your help to launch the campaign. Our research has suggested a new theory that responding with disapproving comments towards those who did not follow the guidelines can be an effective way of educating people about how to behave appropriately. We ought to protect our community and educating non-followers may reinforce them to follow the guidelines. Therefore, you are important in helping us to evaluate whether the disapproving comments can teach non-followers to behave more carefully. In this way, you are helping to keep people safe.

For your information, the public health campaign combating COVID-19 will be

launched in China after it is evaluated. Please help us —— together we can better protect the community at large!

Dr. Wei Zhang
CEO, Chinese Institute of Public Health Research

Ingroup Leader, Leader's Norm and Goal Absent:

National Institute of Public Health Research (US)

Message from CEO

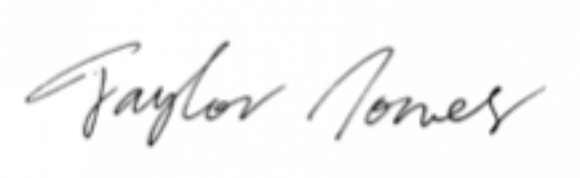
Thank you for showing your interest in helping us evaluate the materials used for a public health campaign about coronavirus.

I am Dr. Taylor Jones, the CEO of a Non-Profit Organization called the National Institute of Public Health Research. We are a charity organization based in the United States. We work on a public health campaign regarding COVID-19.

To protect yourself and others, people need to adhere to the guidelines of keeping a social distance from others, practicing good hygiene such as washing hands correctly, as well as staying at home while sick and following the limits for public gatherings. However, some people are not complying with the guidelines.

This current survey is testing some materials that will be used for a public health campaign. You will be asked to evaluate the materials that will be used in the campaign. In this way, you are helping us to launch the public health campaign.

For your information, the public health campaign combating COVID-19 will be launched in the United States after it is evaluated. Please help us to evaluate the posters!



Dr. Taylor Jones
CEO, National Institute of Public Health Research (US)

Outgroup Leader, Leader's Norm and Goal Absent:

Chinese Institute of Public Health Research

Message from CEO

Thank you for showing your interest in helping us evaluate the materials used for a public health campaign about coronavirus.

I am Dr. Wei Zhang, the CEO of a Non-Profit Organization called the Chinese Institute of Public Health Research. We are a charity organization based in China. We work on a public health campaign regarding COVID-19.

To protect yourself and others, people need to adhere to the guidelines of keeping a social distance from others, practicing good hygiene such as washing hands correctly, as well as staying at home while sick and following the limits for public gatherings. However, some people are not complying with the guidelines.

This current survey is testing some materials that will be used for a public health campaign. You will be asked to evaluate the materials that will be used in the campaign. In this way, you are helping us to launch the public health campaign.

For your information, the public health campaign combating COVID-19 will be launched in China after it is evaluated. Please help us to evaluate the posters!

Dr. Wei Zhang
CEO, Chinese Institute of Public Health Research

Social media posts with only one shaming comment:

Sarah Garrard
Yesterday at 8:55am ·

Removed due to copyright restriction.

Hope this is enough food for my family!

Like · Comment · Share

2 people like this.

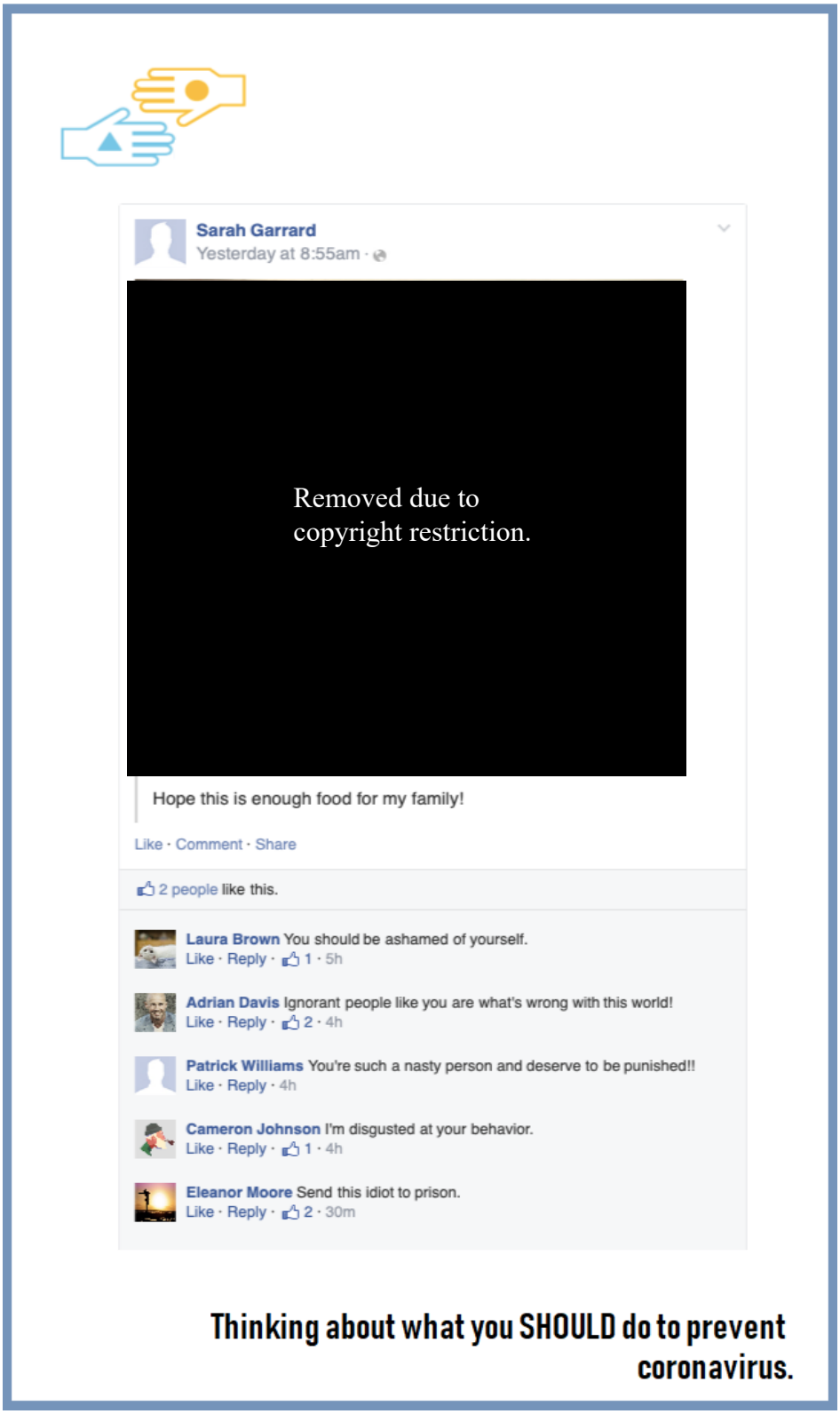
Eleanor Moore Send this idiot to prison.
Like · Reply · 2 · 30m

Thinking about what you SHOULD do to prevent coronavirus.


James Wilson
Yesterday at 3:23pm
Who cares about social distancing when drunk? Happy birthday David Collins Thanks for inviting us to the party!
Like · Comment · Share
4 people like this.
2 shares
Michael Anderson You're a scum and I'm offended by your existence
Like · Reply · 1 · 8h
Thinking about what you SHOULD do to prevent coronavirus.


Robert Stewart
Today at 9:54am
My daughter was coughing last night, but I still sent her to school this morning People really need to calm down. Coughing is NOT a big deal
Like · Comment · Share
1 people like this.
Alan Martin You are a piece of trash!
Like · Reply · 1 · 20m
Thinking about what you SHOULD do to prevent coronavirus.

Social media posts with only one shaming comment:



James Wilson
Yesterday at 3:23pm ·

Who cares about social distancing when drunk? Happy birthday David Collins Thanks for inviting us to the party!

Like · Comment · Share

4 people like this.

2 shares

Ann Lopez Pure TRASH...
Like · Reply · 1 · 10h

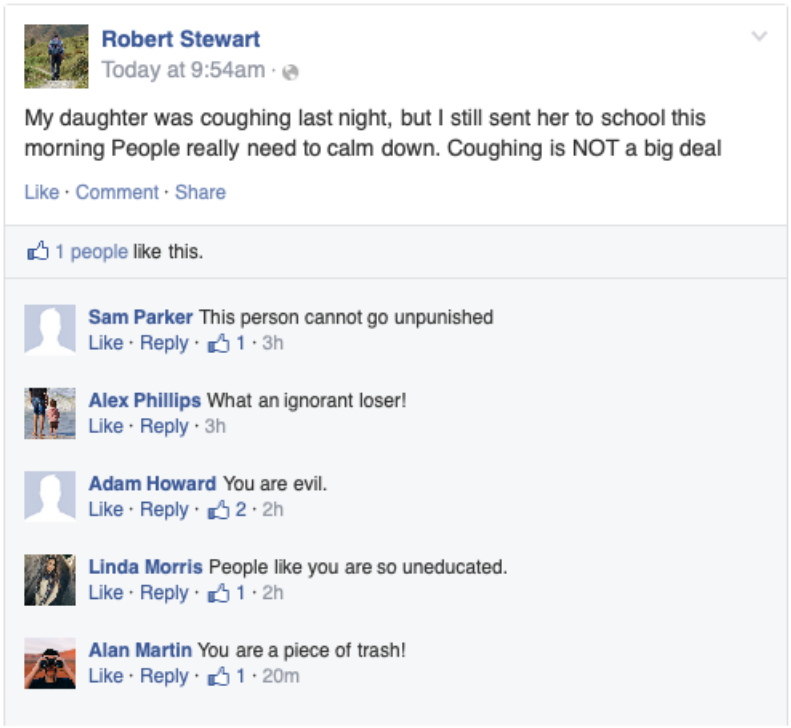
Julie Miller You make me sick
Like · Reply · 2 · 9h

Harry Brown A Real PIG!
Like · Reply · 9h

Peter Nelson Your behavior shows an underdeveloped level of maturity.
Like · Reply · 1 · 8h

Michael Anderson You're a scum and I'm offended by your existence
Like · Reply · 1 · 8h

Thinking about what you SHOULD do to prevent coronavirus.



Thinking about what you SHOULD do to prevent coronavirus.

Study 4

Additional Measures

We measured the following variables but did not provide the details of them in the chapter: Participants were asked to respond to the questions about their evaluations of the situation (perceived threat, certainty, control, personal relevance), emotions, perceived norms on social media, usage of social media and types of platforms, attitudes about American people, Chinese people, and China, knowledge about the pandemic, and their views about the public health campaign. Responses were given on a 7-point scale anchored 1 = *Strongly disagree* to 7 = *Strongly agree*, unless specified.

Perceived Threat. Five items were used to measure people's perceived threat about the pandemic: "Thinking about the coronavirus pandemic, how important to you is each of these concerns?" (Physical health, Mental Health, Supply of daily necessities, Jobs and employment, Income), $\alpha = .76$. Responses were given on a 7-point scale anchored 1 = *Not important at all* to 7 = *Very important*.

Perceived Certainty. Three items were used to measure people's perceived certainty about the situation: "Thinking about the coronavirus pandemic," … "how well do you understand what is happening around you in this situation?", "how uncertain are you about what is happening in this situation?" (reverse coded), "how well can you predict what is going to happen in this situation?", $\alpha = .20$. Responses were given on a 11-point scale anchored 1 = *Not at all* to 11 = *Extremely*.

Perceived Control. Three items were used to measure people's perceived control about the situation: "Thinking about the coronavirus pandemic, to what extent" … "do you feel that circumstances beyond anyone's control is influencing what is happening in this situation?", "do you feel that you have the ability to influence what is happening in this

situation?", "do you feel that someone other than yourself is controlling what is happening in this situation?", $\alpha = .17$. Responses were given on a 11-point scale anchored 1 = *Not at all* to 11 = *Extremely*.

Personal Relevance. Two items were used to measure the personal relevance: "Thinking about the coronavirus pandemic, to what extent" … "do you feel that it is relevant to you?", and "do you feel that it is an important issue to you?", $r = .76$, $p < .001$. Responses were given on a 11-point scale anchored 1 = *Not at all* to 11 = *Extremely*.

Emotions. Two items were used to measure each of the emotions (fear, $r = .63$, $p < .001$; worry, $r = .64$, $p < .001$; outrage, $r = .83$, $p < .001$; anger, $r = .83$, $p < .001$; hope, $r = .86$, $p < .001$; confusion, $r = .88$, $p < .001$): "When I think about the coronavirus pandemic, I feel … (afraid/worry, outraged, angry, hopeful, confused) of the situation for myself.", and "When I think about the coronavirus pandemic, I feel … (afraid/worry, outraged, angry, hopeful, confused) about what my loved ones should do.".

Perceived Norm on Social Media. Four items were used to measure the perceived social norm on social media. Two of them measured the perceived increase in hostility, "In general, I have noticed an increase of hostility on social media since the coronavirus pandemic." and "In general, I have seen more disapproving comments on social media than usual since the coronavirus pandemic.", $r = .84$, $p < .001$. The other two measured the perceived increase in friendly discussions, "In general, I have seen more nice comments on social media than usual since the coronavirus pandemic." and "In general, I have seen an increase of friendly discussions on social media since the coronavirus pandemic.", $r = .74$, $p < .001$.

Usage of Social Media. One open-ended question was used to measure people's daily usage of social media: "On average, how many hours per day did you spend on social media

in the past two weeks?" (M = 3.18, SD = 2.88). Another item was used to measure people's usage of social media since pandemic: "Compared to a typical day before the coronavirus pandemic, how long did you spend on social media per day in the past two weeks?", on a 7-point scale anchored 1 = *Far below average* to 7 = *Far above average* (M = 4.55, SD = 1.27).

Identification with American People. Two items were used to measure identification with the American people, "I feel close to the American people." and "I identify with American people.", r = .79, p < .001.

Identification with Chinese People. Two items were used to measure identification with the American people, "I feel close to the Chinese people." and "I identify with Chinese people.", r = .84, p < .001.

Opinion about China. One item was used to measure people's opinion about China, "What is your opinion of China?", on a 7-point scale anchored 1 = *Strongly unfavourable* to 11 = *Strong favourable*.

Knowledge about Pandemic. Two items were used to measure people's knowledge about the pandemic: "I have a good knowledge about how to protect myself in the coronavirus pandemic." and "I have a good knowledge about what to do in the coronavirus pandemic.", r = .72, p < .001.

Social Media Platforms. One item was used to measure the social media platforms participants use: "Please select the social media platforms you use (you may select more than one)" (response options: Facebook, Twitter, Instagram, Tumblr, Reddit, YouTube, Online Newspapers, Other). Facebook (84.5%) and YouTube (74.4%) were two the most popular social platforms, followed by Instagram (62.6%), Twitter (50.7%), Reddit (40.1%), Online Newspapers (25.6%), Tumblr (4.7%), and other (6.9%).

Comprehension Checks. To ensure the participants paid attention and comprehended our manipulations, they were asked to answer three questions after reading the message from CEO. Three of them were multiple-choice questions: “Where does the CEO come from?” (response options: Australia, Korea, China, US), “Who organized this training program?” (response options: a government department, a university, a non-profit organization, an IT company), and “What is this public campaign about?” (response options: Increase people's awareness of how to maintain a good mental health, Educating people about the importance of getting vaccinated, Increase people's awareness of how to prevent the novel coronavirus infections, Educating people about how to prevent human immunodeficiency virus (HIV) infections. Most participants answered the multiple-choice apprehension checks correctly, with only 15.0% of the participants answered one of the comprehension checks incorrectly. An additional statement, “The public health campaign can help to prevent the spread of the coronavirus disease (COVID-19)”, was used to measure whether participants found the public health campaign effective after reading the letters from the CEO.

View about the Campaign. One open-ended question was used to measure people’s view about the campaign: “Generally, what do you think of this campaign?”.

Norm Manipulation

Two scales were calculated based on the responses of the open-ended question for norm: 1) perceived norm (specific). Participants who mentioned the norm specifically (e.g., we should shame, provide public disapproval, or call out people with comments) scored 1, whereas who did not recall the specific norm were given a score of 0 and 2) perceived norm (non-specific). Participants who mentioned the norm at least vaguely (e.g., to educate or to punish people failed to comply with the guidelines) were given a score of 1, whereas who did not recall the norm even vaguely were given a score of 0.

Table D1

Percentages of participants perceived the leader's norm by conditions (leader's nationality)

Variable	Leader's norm and goal presented[a]		Leader's norm and goal not presented[a]	
	Ingroup[b] (n = 97)	Outgroup[b] (n = 106)	Ingroup (n = 99)	Outgroup (n = 104)
Perceived norm (Yes/No)[c]	92.8%	85.8%	42.4%	34.6%
Perceived norm (specific)[d]	61.9%	62.2%	4.0%	1.0%
Perceived norm (non-specific)[d]	72.2%	70.7%	8.1%	3.8%

Note. [a] The presence of leader's norm and goal was coded as 1 = leader's norm and goal was absent, 2 = leader's norm and goal was present. [b] Nationality of the leader was coded as 1 = ingroup/American, 2 = outgroup/Chinese. [c] Perceived norm (Yes/No) were coded as 1 = Yes, 2 = No. [d] Perceived norm (specific) and perceived norm (non-specific) were coded as 1 = Absent, 2 = Present.

Table D2

Logistic coefficients for predicting perceived norm variables

Predictor	*B* (*SE*)	Wald[a]	*p*	Odds Ratio	95% CI for odds ratio
Outcome Variable: Perceived norm (Yes/No)[b]					
Constant	2.41 (0.50)	23.42	<.001	11.15	NA
Nationality[c]	0.45 (0.25)	3.28	.07	1.56	[0.96, 2.53]
Presence[d]	-2.61 (0.27)	92.98	<.001	0.07	[0.04, 0.13]
Outcome Variable: Perceived norm (specific)[e]					
Constant	-6.37 (1.04)	37.36	<.001	0.002	NA
Nationality[c]	0.13 (0.31)	0.17	.68	1.13	[0.62, 2.06]
Presence[d]	3.51 (0.49)	51.28	<.001	33.36	[12.77, 87.12]
Outcome Variable: Perceived norm (non-specific)[e]					
Constant	-4.89 (0.80)	37.05	<.001	0.01	NA
Nationality[c]	0.06 (0.32)	0.03	.86	1.06	[0.57, 1.99]
Presence[d]	3.10 (0.37)	71.95	<.001	22.11	[10.81, 45.22]

Note. [a] All Wald tests presented in this table has a degree of freedom of 1. [b] Perceived norm (Yes/No) were coded as 1 = Yes, 2 = No. [c] Nationality of the leader was coded as 1 = ingroup/American, 2 = outgroup/Chinese. [d] The presence of leader's norm and goal was coded as 1 = leader's norm and goal was absent, 2 = leader's norm and goal was present. [e] Perceived norm (specific) and perceived norm (non-specific) were coded as 1 = Absent, 2 = Present.

Table D3

Fixed-Effects ANOVA results using identification with leader as the criterion

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	$_{partial}\eta^2$
Leader's group membership (Group)[a]	0.79	1	0.79	0.31	.58	.01
Presence of leader's norm and goal (Presence)[b]	3.53	1	3.53	1.40	.24	<.01
Order of identification with leader (Order)[c]	34.33	1	34.33	13.61	< .001	.07
Number of shaming comments (Comment)[d]	2.21	1	2.21	0.87	.35	<.001
Order × Group	1.58	1	1.58	0.63	.43	<.01
Order × Presence	16.07	1	16.07	6.37	.01	.04
Group × Presence	0.78	1	0.78	0.31	.58	.01
Order × Comment	2.21	1	2.21	0.88	.35	<.01
Group × Comment	0.84	1	0.84	0.33	.56	<.01
Presence × Comment	1.56	1	1.56	0.62	.43	<.001
Order × Group × Presence	1.26	1	1.26	0.50	.48	.01
Order × Group × Comment	0.90	1	0.90	0.36	.55	<.01
Order × Presence × Comment	0.12	1	0.12	0.05	.83	<.001
Group × Presence × Comment	1.76	1	1.76	0.70	.40	<.01
Order × Group × Presence × Comment	0.28	1	0.28	0.11	.74	<.001
Error	983.37	390	2.52			

Note. [a] Leader's group membership was coded as 0 = ingroup/American, 1 = outgroup/Chinese. [b] Presence of leader's norm and goal was coded as 0 = leader's norm

and goal were absent, 1 = leader's norm and goal were present. [c] Order of identification with the leader was coded as 0 = before, and 1 = after. [d] Number of shaming comments was coded as 0 = one shaming comment, 1 = five shaming comments.

Table D4

Comparison between group means based on the order of identification with leader and presence of leader's norm and goal

Comparison between groups (Order × Presence)	M_{diff}	*p*
After: Absent – Before: Absent	-1.47	<.001
Before: Present – Before: Absent	-0.80	<.01
After: Present – Before: Absent	-1.03	<.001
Before: Present – After: Absent	0.66	0.02
After: Present – After: Absent	0.43	0.20
After: Present – Before: Present	-0.23	0.73

Table D5

Fixed-Effects ANOVA results using perceived nobleness as the criterion

Predictor	Sum of Squares	*df*	Mean Square	*F*	*p*	partial η^2
(Intercept)	684.02	390	1.75			
Order of identification with leader (Order)[a]	4.74	1	4.74	2.71	.10	<.01
Presence of leader's norm and goal (Presence)[b]	12.36	1	12.36	7.05	<.01	.01
Leader's group membership (Group)[c]	0.06	1	0.06	0.03	.86	<.001
Number of shaming comments (Comment)[d]	1.34	1	1.34	0.76	.38	<.001

Order × Presence	4.46	1	4.46	2.54	.11	<.01
Order × Group	0.28	1	0.28	0.16	.69	<.001
Presence × Group	1.85	1	1.85	1.06	.30	<.01
Order × Comment	0.01	1	0.01	0.01	.95	<.001
Presence × Comment	0.002	1	0.002	0.002	.97	<.001
Group × Comment	0.03	1	0.03	0.02	.90	<.001
Order × Presence × Group	0.29	1	0.29	0.16	.69	<.001
Order × Presence × Comment	0.61	1	0.61	0.35	.56	<.001
Order × Group × Comment	1.97	1	1.97	1.12	.29	<.01
Presence × Group × Comment	3.74	1	3.74	2.13	.15	<.01
Order × Presence × Group × Comment	0.74	1	0.74	0.42	.52	.001
Error	684.02	390	1.75			

Note. [a] Order of identification with the leader was coded as 0 = before, and 1 = after. [b] Presence of leader's norm and goal was coded as 0 = leader's norm and goal were absent, 1 = leader's norm and goal were present. [c] Leader's group membership was coded as 0 = ingroup/American, 1 = outgroup/Chinese. [d] Number of shaming comments was coded as 0 = one shaming comment, 1 = five shaming comments.

Table D6

*Regression coefficients for the two-way interaction*s *predicting perceived effectiveness of online shaming posts*

Predictor	b (SE_b)	95% CI for b
Outcome variable: Perceived effectiveness, R^2 = .05, MSE = 1.45, $F(4, 401) = 5.28$, $p < .001$		
Constant	2.60**	[2.30, 2.91]
Presence of norm and goal (Presence) [a]	0.08	[-0.34, 0.49]
Order of identification with leader (Order) [b]	-0.05	[-0.45, 0.35]
Presence × Order	0.48	[-0.09, 1.06]
Covariate: No. of shaming comments[c]	0.51**	[0.22, 0.80]

Note. CI = confidence interval. ** indicates $p < .01$.
[a] Presence of leader's norm and goal was coded as 0 = leader's norm and goal were absent, 1 = leader's norm and goal were present. [b] Order of identification with the leader was coded as 0 = before, and 1 = after. [c] Number of shaming comments was coded as 0 = one shaming comment, 1 = five shaming comments.