A study of the emergence, impacts, and responses to trolling in the Australian news media

Delysha Pick
Edith Cowan University

Follow this and additional works at: https://ro.ecu.edu.au/theses

Part of the Arts and Humanities Commons

Readers should be aware that this thesis discusses offensive and discriminatory language and behaviour, sexual violence, and suicide

Recommended Citation

This Thesis is posted at Research Online.
https://ro.ecu.edu.au/theses/2576
A study of the emergence, impacts, and responses to trolling in the Australian news media

Delysha Pick
Student Number: 

Principal Supervisor: Dr Kayt Davies
Associate Supervisor: Professor Trevor Cullen

Doctor of Philosophy

School of Arts and Humanities
Edith Cowan University

December 2021
Copyright statement use of thesis

This copy of the property of Edith Cowan University. However, the literary rights of the author must also be respected. If any passage from this thesis is quoted or closely paraphrased in a paper or written work prepared by the user, the source of the passage must be acknowledged in the work. If the user desires to publish a paper or written work containing passages copied or closely paraphrased from this thesis, which passages would in total constitute an infringing copy for the purpose of the Copyright Act, he or she must first obtain the written permission of the author to do so.
Declaration

I certify that this thesis does not, to the best of my knowledge and belief:

(i) incorporate without acknowledgement any material previously submitted for a degree or diploma in any institution of higher education.
(ii) contain any material previously published or written by another person except where due reference is made in the text; or
(iii) contain any defamatory material

I also grant permission for the Library at Edith Cowan University to make duplicate copies of my thesis as required.

Signed:  

Date:  December 23rd 2021
Acknowledgements

I would sincerely like to thank my principal supervisor Dr Kayt Davies for her support, guidance, and patience throughout my learning journey as I am certainly not a textbook PhD student candidate. Her support has proved invaluable, as has her ability to suggest articles that have enabled me to grasp difficult learning concepts more precisely. I would also like to thank my associate supervisor Professor Trevor Cullen for his guidance with clear and concise thesis structure, proofreading and patience.

Thanks also to Dr John Hall, Dr Kylie Stevenson and Bethany Andersson for their editing and writing style help, and to Neil Ferguson for his qualitative analysis assistance, and to Dr Floyd Holmes for his quantitative analysis support.

I would like to thank the ten journalists who took part in this study, for both their time and honesty in shedding light on such a personal and confronting issue. I thank media legal experts Mark Polden and Professor Mark Pearson, along with academics Dr Jenna Price and Dr Jennifer Beckett, for their willingness to speak with me.

On a personal level, I’d like to thank my parents Ian and Delene Pick for their unconditional love and support, and for using all their financial resources (for their only child) to get me through private school, and building the foundation of my love of knowledge, and search for personal academic achievement to grow.

And to my incredible partner, Sebastian Stall, for the many late nights, the tears, the shoulder massages, the energy drinks, the limited space on the couch between endless piles of paperwork, and my three-hour explanation attempts of my theoretical and epistemological problems in response to a simple question “how are you doing babe?” I could not have achieved this without you.
Abstract

This study analyses the impact of trolling on journalists in the Australian news media between 2015 and 2021, mapping its emergence and the responses of a variety of stakeholders. By 2015, trolling had cemented itself as a complex social issue prevalent in media discourse and has been the subject of a growing body of media and communications research over two decades.

A grounded theory approach, informed by feminist critical theory, functionalism and Ackoff’s (1974) Systems Theory, was used to investigate the impact of trolling on victims, and responses to trolling by anti-trolling campaigners, Twitter, Facebook, newsrooms, and the legal profession. Ackoff’s theory suggests that social problems cannot be resolved by considering stakeholders in isolation, and that a multi-site or multi-level approach to a problem is more likely to succeed. The research process included interviews with 10 trolled journalists, and content analysis of a range of archival evidence relevant to the other stakeholders, such as news articles, statements by social media companies, and Australian legislation. At points where professional insight into the results of the content analysis was required, the research included interviews with relevant experts.

The results of the research include identification of types of trolling frequently encountered by journalists; documentation of widespread frustration with Twitter’s inconsistent enforcement of policies that ban abusive content on its service; and acknowledgement of Facebook’s attempts to address multiple stakeholders involved in trolling and cyber-bullying. Newsrooms’ responses were categorised into ten themes that address the needs of various stakeholders, and an analysis of the relevant current Australia legislation found that while a range of statutes can be drawn upon, their use in relation to trolling has been scarce to date, with experts suggesting that more education of the public and law enforcement officers would enhance legal protection.

This study concludes that trolling is a wicked problem, meaning it is complex, dynamic, and difficult to navigate; thus, resolution strategies should involve collaborative approaches by multiple key stakeholders.
Contents

Title page i
Copyright statement ii
Declaration iii
Acknowledgements iv
Abstract v
Contents vi
List of figures ix

1. Introduction 1
   1.1 Overview 1
   1.2 Trolling: a key term 2
   1.3 Epistemology 2
   1.4 Theoretical framework 3
      1.4.1 Feminist critical theory 4
      1.4.2 Ackoff’s systems theory 4
      1.4.3 Functionalist theory 6
   1.5 Rationale 6
   1.6 Aims and significance 7

2. Literature review 8
   2.1 Feminist context of the research 8
      2.1.1 Rape culture 10
   2.2 Systems theory context of the research 12
      2.2.1 Ackoff’s systems approach 13
      2.2.2 Wicked problems 15
      2.2.3 Functionalist paradigm 16
   2.3 Concept of trolling as research focus 17
      2.3.1 Evolution of definition and understanding 17
      2.3.2 Evolution of use in the Australian news media 18
      2.3.3 Evolution of academic literature on trolling 18
   2.4 New media and trolling 21
      2.4.1 News transition from print to online 21
      2.4.2 Interactivity and trolling 22
      2.4.3 Anonymity and trolling 24
   2.5 Trolling victims 24
      2.5.1 Trolled journalists 24
      2.5.2 Trolling and gender 26
3. Methodology

3.1 Over-arching research question

3.1.1 Research sub-questions

3.2 Research design

3.2.1 Methodological approach

3.2.1.1 Emergence

3.2.1.2 Impacts

3.2.1.3 Responses

3.2.1.3.1 Campaign responses

3.2.1.3.2 Twitter responses

3.2.1.3.3 Facebook responses

3.2.1.3.4 Newsrooms responses

3.2.1.3.5 Legal responses

3.3 Ethical considerations

3.4 Limitations

4. Findings

4.1 Emergence results and discussion

4.1.1 Quantitative data results

4.1.2 Quantitative analysis

4.1.3 Qualitative data results

4.1.4 Qualitative thematic analysis

4.2 Impacts results and discussion

4.2.1 In-depth interview results

4.2.3 Thematic analysis

4.3 Campaign response results and discussion

4.3.1 Qualitative data results

4.3.2 Qualitative and quantitative thematic analysis

4.4 Twitter response results and discussion

4.4.1 Qualitative data results

4.4.2 Thematic critical analysis

4.5 Facebook social media response results and discussion

4.5.1 Facebook support strategies

4.5.2 Facebook feature strategies

4.5.3 Facebook functionality strategies

4.5.4 Facebook research strategies

Qualitative data results and thematic analysis
## List of figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>The trolling paradigm</td>
<td>4</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Stakeholders from the Trolling Paradigm categorised by levels in Ackoff's (1974) Systems Model</td>
<td>5</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Frequency of trolling articles in the Australian news media</td>
<td>40</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Frequency of trolling related news articles in the thematic content analysis of this study</td>
<td>41</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Gender comparison of trolling victims in the thematic content analysis of this study</td>
<td>43</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Comparison of victim gender and frequency of rape threats in the thematic content analysis of this study</td>
<td>44</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Comparison of victim gender and frequency of death threats in the thematic content analysis of this study</td>
<td>45</td>
</tr>
<tr>
<td>Figure 8</td>
<td>A significant portion of the 118 victims with redacted gender identities are likely to be female</td>
<td>47</td>
</tr>
<tr>
<td>Figure 9</td>
<td>A chronological timeline of the evolution of trolling in the Australian news media derived from chapter 4.1 data</td>
<td>51</td>
</tr>
<tr>
<td>Figure 10</td>
<td>Identified emerging themes linked with theoretical framework</td>
<td>66</td>
</tr>
<tr>
<td>Figure 11</td>
<td>Frequency of trolling comments received by participants</td>
<td>67</td>
</tr>
<tr>
<td>Figure 12</td>
<td>Clustered trolling comments received by participants</td>
<td>67</td>
</tr>
<tr>
<td>Figure 13</td>
<td>Specific trolling content received by participants</td>
<td>68</td>
</tr>
<tr>
<td>Figure 14</td>
<td>Platforms participants received trolling content on</td>
<td>69</td>
</tr>
<tr>
<td>Figure 15</td>
<td>Identified emerging themes within Ackoff's systems model</td>
<td>78</td>
</tr>
<tr>
<td>Figure 16</td>
<td>A chronological timeline of anti-trolling campaigns in the Australian news media derived from chapter 4.3 data</td>
<td>84</td>
</tr>
<tr>
<td>Figure 17</td>
<td>The cumulative effect of identified trolling campaigns on shaping universally accepted social tolerances of workplace sexual harassment derived from chapter 4.3 data</td>
<td>89</td>
</tr>
<tr>
<td>Figure 18</td>
<td>A snapshot of the racism-based anti-trolling campaigns identified in 2015</td>
<td>93</td>
</tr>
<tr>
<td>Figure 19</td>
<td>Twitter's quality filter feature for users</td>
<td>108</td>
</tr>
<tr>
<td>Figure 20</td>
<td>A timeline of Twitter's first response to trolling</td>
<td>112</td>
</tr>
<tr>
<td>Figure 21</td>
<td>A timeline snapshot of the trolling of Zelda Williams</td>
<td>119</td>
</tr>
<tr>
<td>Figure 22</td>
<td>Twitter's previous nine-part questionnaire for reporting abuse</td>
<td>124</td>
</tr>
<tr>
<td>Figure 23</td>
<td>Twitter's block feature introduced in 2014</td>
<td>125</td>
</tr>
<tr>
<td>Figure 24</td>
<td>The state of internet freedom around the world in 2019</td>
<td>146</td>
</tr>
<tr>
<td>Figure 25</td>
<td>The initial interface of Facebook's Safety Centre when launched in 2010</td>
<td>150</td>
</tr>
<tr>
<td>Figure 26</td>
<td>The most recent 2020 interface of Facebook's Safety Centre</td>
<td>151</td>
</tr>
<tr>
<td>Figure 27</td>
<td>A table to show the study's tiered thematic analysis of Facebook feature responses to trolling derived from chapter 4.5 data</td>
<td>165</td>
</tr>
</tbody>
</table>
Figure 52  Australian criminal laws for sexual vilification that trolled journalists could use derived from chapter 4.7 data 331
Figure 53  Australian criminal laws for gender vilification that trolled journalists could use derived from chapter 4.7 data 333
Figure 54  Australian criminal laws for HIV/AIDS vilification that trolled journalists could use derived from chapter 4.7 data 335
Figure 55  Australian civil laws for past economic loss that trolled journalists could use derived from chapter 4.7 data 337
Figure 56  Australian civil laws for future economic loss that trolled journalists could use derived from chapter 4.7 data 339
Figure 57  Australian civil laws for superannuation loss that trolled journalists could use derived from chapter 4.7 data 340
Figure 58  Australian civil laws for non-economic loss that trolled journalists could use derived from chapter 4.7 data 341
Figure 59  Australian civil laws for duty of care that trolled journalists could use 343
Figure 60  Australian civil laws for workers compensation that trolled journalists could use 349
Figure 61  Australian civil laws for defamation that trolled journalists could use 351
Chapter 1

Introduction

Journalists have always suffered abuse; however, the physical, psychological, professional, and economic effects of trolling illustrate the severity of the problem. Abuse and harassment have become commonplace for many Australian journalists, not only because of the interactive function of Web 2.0, but as a broader reflection of current politics, economics, and ideologies. Women are disproportionately targeted by trolls, with more frequent and aggressive attacks. Using grounded theory, this study adopts a systems approach, informed by feminist critical theory and functionalism, to address each stakeholder group: trolled journalists, anti-trolling campaigns, social media organisations, newsrooms, and legal.

Trolling is a complex social problem because it involved the actions of many people, on many people, and is intertwined with broader issues including freedom of the press, freedom of speech, rape culture, and workplace health and safety. The rise of participatory journalism mechanisms, including social media, have enabled audiences to speak back to journalists with greater intensity and anonymity than ever before. The trolling phenomenon is tangled up with speech rights and increasing public critiques of social institutions, like the media, as well as rising tides of disinformation. In 2022, the digital abuse and harassment of journalists is an enduring problem that requires further systematic research to determine the true scope, scale, and potential resolutions to the problem.

1.1 Overview

This research, a study of the emergence and impacts of and responses to trolling in the Australian news media, was conducted over six years from 2015 to 2021, and provides the first comprehensive analysis of ‘trolling’ abuse and harassment of journalists in the Australian news media industry, mapping its evolution,
impacts, and industry responses. This mixed-methods study uses the grounded theory approach (described in the chapter 3 methodology section) to code both qualitative and quantitative data, from which developing patterns emerge for further thematic analysis. Data collection methods of in-depth interviews, archival evidence and thematic content analysis were included. The timing of this study was fortuitous, given the expeditious growth of technological innovation and subsequent economic restructures of journalism globally over the period of this study. By the conclusion of the study, trolling had cemented itself as a complex social issue prevalent in workplace dialogue, with growing academic research yet no universally recognised resolution.

1.2 Trolling: a key term
While the term ‘troll’ is widely used throughout Australian news media (Binns, 2017; Bossio & Holton, 2021; Gardiner, 2018; Lewis, Zamith & Coddington, 2020; Martin & Murrell, 2020; Martin & Murrell, 2021), an academic understanding of the term was still evolving, resulting in various definitions of its meaning (Phillips, 2015; Tyler, 2012). While definitions of trolling have evolved in the last seven years, when this study commenced in 2015, less had been written and the chief concern of people writing about the social problem was distinguishing trolling behaviour from cyber-bullying and other newsroom interactivity. As Golf-Papez and Veer wrote in 2017, at that point “there had been little discussion about what trolling actually is” (p. 1). Consequently, the definition selected for this study in 2015 was Johnston’s (2014) definition: “trolling is a deliberate attempt to upset or anger someone online” (para. 4) as this was in accord with the context given by West (2015) and Ford (2015). This definition is also similar to the one Golf-Papez and Veer (2017) used, which is that “trolling involved deliberate, deceptive and mischievous attempts to provoke reactions from other online users” (p. 1).

1.3 Epistemology
This applied research study integrates epistemological theory to examine trolling in social contexts, an approach that lays the foundation from which grounded theory emerges (Weerakkody, 2015). The epistemology, or theory of knowledge (Crotty 1998; Hamlyn, 1995; Maynard, 1994) this study is applied is constructivist
theory to better understand the phenomenon of trolling. Constructionism posits that “no objective reality or truth” exists and “reality is socially constructed” (Weerakkody, 2015, p. 10). This research following Phillips (2015) investigates the social construction of trolling. The way trolling has been socially constructed has evolved and been tracked by this study that began with the researcher’s emersion in the preliminary data set from which the statistical analysis emerges.

1.4 Theoretical framework

As this study describes the complex social issue of trolling in various digital spaces of news media, multiple paradigms have been adopted to form a cohesive theoretical framework for the study. This theoretical approach is comprised of three key paradigms: feminist critical theory, Ackoff’s (1974) systems theory, and functionalism. These three key theoretical frameworks underpin the research and are used to examine and investigate various groups of stakeholders affected by trolling. Each of these three key paradigms are briefly introduced below, and further clarified in the literature review. In Figure 1, feminist critical theory is shown to relate to victim and campaign responses; functionalism to relate to social media platforms, newsrooms, and legal responses; and Ackoff’s systems theory to relate to all stakeholder groups. The research theoretical framework, which this study calls the trolling paradigm, is depicted below in the following diagram:
1.4.1 Feminist critical theory

The first theoretical framework underpinning this research process is the feminist critical paradigm and was selected as women are more often and more aggressively trolled than men. Critical theory examines power relations within social phenomenon to empower those adversely affected by patriarchy and racism (Horkheimer, 1972), while feminist theory critiques and resists oppressive social power relations pertaining to gender (Allen, 1999). Adopting a feminist critical theoretical perspective, this study examines patriarchal ideologies in relation to trolling to identify sexism and racism embedded in trolling content and behaviour, which is often aimed at female journalists (Binns, 2017; Gardiner, 2018; Lewis, Zamith & Coddington, 2020; Martin & Murrell, 2020; Martin & Murrell, 2021).

1.4.2 Ackoff's systems theory

The second theoretical framework underpinning this research process is Ackoff's (1974) systems paradigm. Ackoff's systems theory sees society as a system where individual parts must work together in synergy to ensure the system runs efficiently. In order to provide a comprehensive understanding of the social

Figure 1

*The Trolling Paradigm*
problem of trolling, the study will examine each individual element as subsections in order to have a comprehensive understanding of the systemic complexity of trolling.

Figure 2

*Stakeholders from the Trolling Paradigm categorised by levels in Ackoff’s (1974) Systems Model*

For brevity this framework will be referred to throughout the study as the systems model. At Ackoff’s (1974) “self-control” level (p. 20), or the study’s coined individual level, are the victims; the trolled journalists, and the perpetrators of the problem; the trolls. At Ackoff’s (1974) “humanisation” level (p. 34), or the study's coined support, peer, and community level, are groups which are able to offer support to the individuals experiencing the problem which include newsrooms, family, peers, community and campaigns. At Ackoff’s (1974) “environmentalisation” level (p. 54), or the study's coined governance level, is
legal and, increasingly more so, social media platforms. At the highest level, social institutions have the power to implement policies, regulations, and/or laws that inhibit trolling behaviours. Each chapter of the study addresses an individual stakeholder, and the discussion section explores the systematic interactions and interdependencies between stakeholders. Analysis of relationships between stakeholder groups is critical to Ackoff’s (1974) System Theory that suggests a holistic approach that works on the problem at all levels is needed to address social problems such as trolling.

1.4.3 Functionalist theory
The third theoretical framework, complementary to Ackoff’s systems theory, is the functionalist paradigm. Functionalism is a theory of mind, here used in sociological terms to explain the behaviour of those trolling journalists and argues that a perpetrator’s behaviour is directly related to the consequence of that behaviour (Spencer, 1893; Parsons, 1977; Macionis 1995). When applied to trolling, functionalism, with its focus on how parts of a social system interact, provides a lens through which to examine online trolling behaviour. Such a perspective provides grounds for the study’s argument that the anonymity afforded by the internet presents a lack of consequence for the perpetrators, and therefore contributes to the increased frequency and extent of explicit trolling content.

1.5 Rationale
This study was inspired by the first trolled journalists to speak out. In 2015, feminist writer for the US publication Jezebel Lindy West and Australian feminist writer and regular columnist for The Daily Telegraph Clementine Ford, were among the first to articulate that trolling was a social problem specifically effecting working journalists. While some literature argues that trolling is a natural consequence of freedom of speech (Thompson, 2019), others have reported that ‘don’t feed the trolls’ is a frequently suggested coping strategy (March, 2016). A study by Martin and Murrell (2020) found that trolled journalists were expected to develop resilience on the job. However, West (2015) and Ford (2015) had criticised reliance on personal resilience of journalists and the adage ‘don’t feed
the trolls’ suggesting these approaches gravely underestimated the depth of the problem. The alarm raised by West and Ford indicated a need for action on the part of people other than the trolled journalists themselves. Ackoff’s system theory framework provides a pathway to determining who the other stakeholders are and analysis using this framework can determine what actions need to be taken in order alleviate the problem. To access the extent of the impact of trolling on Australian journalists, research needed to be conducted to determine the scale and scope of the problem. And so, this became the first stage of my research project. As the alarm had been raised by journalists working in newsrooms, a grounded approach was deemed necessary to be sure to capture the lived experience of journalists encountering trolling. Guided by the Ackoff framework, the study looks at a range of stakeholders and the actions with regards to trolling. This would enable the research to identify gaps that are sites of potential action on the part of stakeholders of various levels of the framework.

1.6 Aims and significance
The study aimed to examine how the trolling of Australian journalists (and its characterisation as a social problem) emerged as a topic of social discourse; to investigate the impacts on Australian journalists and whether these impacts were gendered; to document responses by other stakeholders involved with the social problem. This study is significant in that it is the first systemic data collection and analysis of trolling in the Australian news media. The intention was to contribute to the development of practical resolutions to the problem of trolling based on academically rigorous evidence. In a multilayered approach, this study has canvased opinions about best-practice journalism in the workplace and suggests specific implementation strategies that Australian journalism educators could employ in their curricula, addressing the lack of content currently available. In doing so, this study contributes to the preparation of journalism students for a workplace in which trolling is prevalent.
Chapter 2

Literature Review

This literature review will expand on the three key paradigms that make up the Trolling Paradigm. Subheading will include the feminist context of the research, rape culture, the systems theory context of research, Ackoff’s Systems approach, and the functionalist paradigm. The section will also explore the concept of trolling as a research focus through explanation of the evolution of definition and understanding, evolution of use in the Australian news media, and gaps in academic literature. While digital media trolling is explored through discussions about news transition from print to online, interactivity and anonymity, trolling victims are also looked at through exploration of both trolled journalists and Australian female trolled journalists specifically. As this study is based on grounded theory (as detailed in the next chapter) which requires cycles of data gathering and theoretical development, each stage brought to light new issues. For example, the issue of racism was raised in the investigation of anti-trolling campaigns. Literature about these supplementary issues is explored in context in the relevant chapters. As a result, this literature review is limited to exploration of issues that were evident at the outset of the research.

2.1 Feminist context of the research

Feminist scholars have had a focus on critiques of power and violence, and more recently on the problems of intersectionality and social and cultural exclusion. In order to provide context for this study, it is important to provide a brief historical overview of feminist theory. The first two decades of the 2000s have been a time of exciting, fast-moving developments in feminist politics (Lewis, 2018). As outlined by Munro (2013), the suffrage endured by first-wave feminist leaders in the early twentieth century is widely documented throughout academic literature, addressing property ownership and the right to vote (para. 2). Second-wave feminism in the early 1960s gradually turned its attention to women’s equality in
wider society, coining the phrase ‘the personal is political’ and highlighting the impact of sexism and patriarchy on the private lives of women, further breaking down gender stereotypes (para. 2). Second-wave feminism was often criticised because it appeared to assume that women were one homogenous (white, middle-class) group, sidelining women of colour (Blackwell, 2011; Butler, 2013; Munro, 2013). In the early 1990s third-wave feminism was born, heavily influenced by academic investigations of queer theory, which suggested gender and sexuality are socially constructed. However, third-wave feminism has been criticised for its individual emancipation focus (para. 2). Most recently, commentators argue that the internet itself has enabled a new shift into fourth-wave feminism. Judy Wajcman (2004) is one of a number of feminist scholars suggesting that technologies are gendered in both their design and use. As study by Fichman and Sanfilippo (2014) found that “men and women react differently to online trolling and their perceptions of the impact of trolling on online communities vary” (p. 1). Therefore, from this perspective, it makes sense to use a feminist lens that considers potential differences between genders when investigating trolling of journalists.

Aune and Refern (2013) suggest some feminists point their fingers at ‘the media’ “as if they are a static body of white men sitting in a Hollywood mansion who lock the doors when they see women coming down the path” (p. 171) but in reality, feminism is complex, and patriarchy is a “wicked problem” (Ritchey, 2013, p. 1). Munro (2013) suggests the internet has created a “call-out culture” (p. 23) in which sexism and misogyny can be challenged, a continuing influence of third-wave feminism. Lewis (2018) critiques the idea of “waves of feminism”, suggesting “woven tapestry” (p. 23) as a more accurate description of feminism, representing the interconnectivity of ideas drawn upon over a century of feminism in conjunction with post-structuralist, post-feminist and post-colonialist theories. Lewis argues that the concept of ‘waves of feminism’ depicts each “new wave [as] a homogenous whole, washing away the old, displacing the outdated wave; this does not do justice to the rich variety of feminist thought” (p. 23). Despite widespread critical analysis about its exact categorisation, fourth-wave feminism shares many of the concerns and diverse feminist perspectives developed over the
last century. This thesis will adopt the term contemporary feminism to reflect feminism’s many waves. Contemporary feminism is not only about layers of complexity – but the need to recognise intersectionality, and the different harms that trolling can do to women – and people – from non-Anglo, non-CIS backgrounds.

Although Redfern and Aune (2013) highlighted that some news publications regularly deem feminism to be “dead” (p. 1), Lewis (2018) disputes this idea and argues “contemporary feminist activists continue to address a range of enduring inequalities” (p. 31). Redfern and Aune (2013) suggested the feminist principles of “equality, fairness and non-discrimination” (p. 5) are adopted by younger Australians today as part of socially accepted, cultural norms. The women’s movement of the twentieth century has much to teach those interested in combating gender inequality and female-targeted harassment in this century, such as trolling of woman in the media. This thesis adopts a feminist paradigm to analyse discrepancies in the frequency and severity of trolling in comparison to victim gender. While similar differences may emerge from studies that investigate other characteristics such as race, ethnicity, age, geographic location and prominence, this level on analysis was beyond the scope of this research. Together, education and law reform can help combat destructive social attitudes (Binns, 2017; Gardiner, 2018; Lewis, Zamith & Coddington, 2020; Martin & Murrell, 2020; Martin & Murrell, 2021) and this research will help identify gaps in responses to date.

2.1.1 Rape culture

One of the missions of second-wave feminism was to end the blanket of silence shrouding rape and to bring it to public attention. In Lindy West’s (2015) Jezebel column, she said for years she had endured rape threats. Many of these were specifically in response to columns about rape culture in the routines of some comedians. Rape culture is defined by Wright (2015) as an ethos that “exists in a society or environment in which common social beliefs, attitudes and morals normalise sexual violence, encourage people to associate sex with violence, and minimise the seriousness of sexual abuse” (p. 5). Nearly thirty years after Susan
Brownmiller (1975) brought rape to international attention with her pioneering treatise *Against our will: men, women and rape*, society is faced with the continuing proliferation of glamorised images of rape in the public domain. From the eighteenth-century onwards, rape has been perceived as a crime “in its nature commonly secret” (Ferguson, 1987, p. 91). Employing second-wave feminism rhetoric, Griffin (1981) said it was time to break the “conspiracy of silence” (p. 4) about the real impact of rape. This is what West (2015) was doing in her columns and her commentary about rape jokes. Therefore, consideration of rape culture and its roots provides context for understanding the motivations and behaviour of trolls and the reactions of victims. This study sought to consider the impact of rape culture on the lived experience of trolled Australian journalists.

Discussions of literary and filmic depictions of rape have been particularly prone to pivot on the question of whether such graphic depiction of sexual violence is necessary. Horeck (2013) suggests the representation of rape is perhaps because of the “high emotional and political stakes involved in reading and watching rape in contemporary culture” (p. 8), particularly in the wake of feminist consciousness-raising on the subject, forcing a new understanding of the interrelationship between the imaginary and the real (p. 6). Despite cultural tolerance of rape in virtual fantasy such as gaming, virtual rape culture exposed itself in the real-world in 2006. An example of the transformation of the virtual into the real occurred when an Australian group of teenage boys filmed the sexual assault of a teenage girl. The perpetrators used the footage to produce a DVD distributed to suburban Melbourne schools for $5 and later on Internet sites for up to $60, with excerpts also made freely available on YouTube (Cunningham, 2006). The video recording shows the young men urinating on the girl, setting her hair on fire, throwing her clothes into a river and forcing her to participate in sex acts, with the credits listing the ‘actors’ involved. Eight of the youths were charged in the Melbourne Children’s Magistrates Court in 2007 with assault, manufacturing child pornography, and procuring sexual penetration by intimidation. At the trial and sentencing of the young men responsible, the victim said she was terrified she would be recognised in public and that her life had been changed forever (Medew, 2007). This notable case caused widespread shock and
outrage and marked the start of a growing public awareness of, and scholarly attention to, the ways in which sexual violence is used in the context of digital technologies in Australia. Ten years on, according to Powell and Henry (2017), the combination of accessibility to the internet through smartphone popularity and widespread participation in online social networks has provided a platform for the perpetration of sexual harassment, abuse, and violence. Powell and Henry further suggest it is “unsurprising” (p. 2) that digital technologies might also be used as tools to facilitate sexual-based harms, and that sex-based trolling in the Australian news media industry is often a “daily occurrence” (p. 2).

Rape culture and its manifestation as sexual abuse online is a subject that needs to be approached from an interdisciplinary standpoint (Horeck, 2013) and examined in an array of fields and disciplines. Adopting an interdisciplinary approach, this study draws upon Ackoff’s (1974) systems approach and feminist theory. Following the lead of Wajcman (2004) this study fuses feminism and constructivist theory. The social constructivist approach views technology as being shaped by human action, and interaction and human action as being reciprocally shaped by technology (Hughes, 1986; Pinch & Bijker, 1984). Wajcman (2004) adds that this human action is gendered as are its results. To provide clarity about how gendered online abuse has manifested in a grounded context, this study maps the evolution, impacts of, and responses to, trolling as a social issue. However, gendered online abuse is only one aspect of the trolling of journalists (others include political, religious, and economic) and is a manifestation of larger social inequalities. The impacts of trolling can, and often does, silence female journalists, further diminishing gender equity and freedom of speech (Ireton & Posetti, 2018).

2.2 Systems theory context of the research

Applied research seeks to solve social problems, and Weerakkody (2015) suggests this often relates to the professional practice of journalism, media production and policy making (p. 314) In addition, Heath and Bryant (2000) regard theories as maps that societies use to better understand a given problem, which can
sometimes contribute to a more comprehensive explanation, improvements and possible future resolution. They further suggest that without research findings, theory is mere speculation (p. 5). This applied research study integrates systems theory throughout the research process so it can more systematically explore the actors, interactions and interdependencies that might indicate effective responses to trolling of journalists.

2.2.1 Ackoff’s systems approach

Organisational theorist Russell Ackoff’s (1974) systems approach to social problems that he called “messes” (p. 21) evolved through observation of how society has handled problems, such as crime and poverty, over time. Looking back through history, Ackoff described the Middle Ages (the 5th to the late 15th centuries) as focussed on survival with minimal action taken towards solving social problems. The Machine Age (early-to-mid 20th century) was described as an age in which problems were broken down into parts that were addressed individually, without consideration of their interconnectivity. He was critical of this style of thinking and pointed to problems that arose from failures to consider issues holistically. In today's parlance the term ‘victim blaming’, which victimology researchers Cramer et al (2013) define as the “attribution of blame toward victims” (p. 1), would refer to this type of fragmented approach.

He then coined the term the “Systems Age” (1974, p. 228) to describe the time from the mid 20th century to 1974 and beyond. The Systems Age he proposed was based on the doctrine of “expansionism” (p. 12) that claims all objects, events, and experiences are parts of larger wholes. It turns attention from fragmented elements to wholes with interrelated parts and levels. This systems approach, described by Ackoff, underpins this study that explores how the social problem of trolling, in the context of journalism, manifests and can be tackled on many levels.

Ackoff (1974) suggested that social problems operate on different levels and claimed resolution attempts failed when levels were targeted exclusively. He suggested social problems can be more effectively addressed when the stakeholders are categorised into purposeful groups. These groups can then be
arranged into levels based on their proximity to the harm arising from the problem, and their power to take action. While acknowledging that social problems will always arise and that some old problems will continue to resurface, this thesis adopts the principle that action based on understanding the interconnected nature of social problems is the key to best-practice management and harm minimisation.

He arranges the purposeful groups of stakeholders into a three-level hierarchical system: self-control (p. 20); humanisation (p. 34); and ‘environmentalisation’ (p. 54). The self-control level refers to the individual experiencing the problem and the extent to which those people can resolve the problem by taking action themselves. The humanisation level refers to the support or social group in close proximity to the individual experiencing the problem (victim or perpetrator). This level can include family, friends, colleagues, peers, and supporting community groups. The environmentalisation level refers to social institutions with power to make rules, regulations and laws that inhibit certain behaviours. These levels are illustrated as they are applied to this study in Stakeholders from the Trolling Paradigm categorised by levels in Ackoff’s (1974) Systems Model in Figure 2 (on p. 5). Although there has been some disparate discussion of tackling the problem of trolling at each of these levels individually (Martin, 2018; Dineva and Breitsohl, 2021), no studies as of yet have analysed trolling using a systems approach that might form the foundation of a more coordinated and synthesised response, as this study intends.

While Ackoff’s theory may appear dated and inappropriate for use in a contemporary thesis about rapidly evolving new media, its multilayered approach still has relatively current practical research applications (Badal, 2006; Bernstein, 2011; Eriksson, 2004; Houghton & Metcalfe, 2010; Koller, 2013; Lumbo, 2007; Nuzzaci, 2010). Ackoff’s approach is relevant to this topic because of the need for a systemic analysis to identify sites of potential action. While stakeholder analysis is widely used in media studies, Ackoff’s layered model was considered more appropriate because the vast differences between the various stakeholders, which range from multi-national corporations to individual freelance journalists.

*Chapter 2 Literature review*
Therefore, Ackoff’s widely used systems approach for analysing social problems is a relevant and practical conceptual framework for this thesis.

### 2.2.2 Wicked problems

Trolling presents as a wicked problem that interwines conflicting issues such as press freedom and free speech, that makes effective solutions difficult to navigate. More recently the term wicked problems have been used to describe social problems that are unstable, evolving and difficult to solve (Camillus, 2008). Many researchers (Vernon, 2012; Beutler, 2021) have linked Ackoff’s description of complex problems as ‘messes’ with Rittel’s (1972) description of these problems as ‘wicked’. Beutler (2021) cited both Ackoff and Rittel in describing messes/wicked problems as:

> Complex sets of problems in which many different potential issues are intertwined or linked. They defy normal problem solving. Wicked problems involve incomplete or contradictory knowledge, differing values, multiple assessments of the situation, and a range of stakeholders with relationships among them.

(p. 1)

Phillips (2015) and Rodrigueux (2015) recognised “online abuse isn’t a technological problem; it’s a social problem that just happens to be powered by technology” (para. 17). More specifically, trolling has also been described as a social problem (Bartlett, 2013; Poon, 2008; Macfadyen, 2014) and a wicked one at that (Charman-Anderson in Funnell, 2014), because it involved the actions of many people, on many people, and is intertwined with broader issues including freedom of press, freedom of speech, rape culture, and workplace health and safety. This indicates that a systems approach research framework is justified for this study.

Raising awareness is a key strategy in addressing wicked problems such as trolling, which are complex, dynamic, and difficult to navigate. Wicked problems go beyond the capacity of any one organisation to understand and respond to and working across agency boundaries is increasingly important in tackling them (Australian Public Service Commission, 2007). Changing the behaviour of large groups of people requires a collaborative approach by multiple stakeholder
groups and poses challenges for traditional approaches, given the fluctuating
dynamics of the online environment (Australian Public Service Commission,
2007). The range of traditional levers used to influence behaviour, including
legislation, fines, taxes, other sanctions are often part of the solution, but these
alone may not be sufficient. More innovative, personalised approaches are likely
to be necessary to motivate individuals to actively cooperate in achieving
sustained behavioural change. Wicked problems are frequently unstable,
constantly evolving, and have no clear solution (Camillus, 2008); therefore,
developing effective ways to tackle them is an evolving art.

2.2.3 **Functionalist paradigm**
Initially considered as a systems level of analysis, the functionalist paradigm is
now considered a stand-alone paradigm. Early critics Carey and Kreiling (1974)
argue functionalism was ensnared by logical pitfalls; however, its adoption within
scholarly media research in the twenty-first century (Azeem, 2017; Gregerson,
2017; Teklu, 2014) suggests the theory is an elementary concept, espoused to
intelligibly substantiate findings. In further support, Weerakkody (2015)
concludes functionalism is often applied to research integrating both social and
media studies, enabling the researcher to effectively access how and why social
issues such as trolling arise and what response strategies should be improved or
implemented.

Crucially in sociological research, functionalism argues that the reason people act
in a specific way is a direct result of the known consequence of that given
behaviour. This is the basis of the uses and gratifications theory (Blumler &
Katz, 1974), which sees an “active audience” (Blumler, 1979, p. 13) consuming media
messages for both its associated functions (uses) and pleasures (gratifications),
replacing the preconceived image of the audience as “passive victims” of the media
(p. 10). Blumler argues the uses and gratifications approach of the early 1960s
revitalised the measurement of media effects that was “too often voiced in a sort
of theoretical vacuum” (p. 15) and could not foresee the development of the two-
way communication model of the internet. Functionalism stems suggests users
engage with escapist media content for pleasure, such as reality television,
romance, and sport; however, in the case of trolling this thesis will argue that pleasure is also linked to the intentional harm of others. Weerakkody (2015) says, “we may do something unfair towards others because we know it hurts them, which is pleasurable to us” and further adds “some people may be more willing to hurt others if they were assured of no negative repercussions” (p. 30). The anonymity enabled by the internet provides a known lack of consequences (Suler, 2004), and a functionalist reading of trolling would suggest that trolls may be more willing to hurt others because of the known absence of negative repercussions, thus contributing to the frequency and extent of explicit and vulgar content.

2.3 Concept of trolling as research focus

2.3.1 Evolution of definition and understanding

As the phenomena of the trolling of journalists emerged, so too did attempts to describe and define it. The first definitions were colloquial ones. By 2015, journalists, such as Johnston (2014) had formulated a definition based on emerging understanding of the term, and by 2017, researchers such as Golf-Papez and Veer had formulated one of the early academic definitions. Since then, academic definitions have evolved. By 2021, Martin and Murrell discuss trolling as a challenging form of “dialogic journalism” (p. 845) which they propose “as a future framework for analysing news conversation’s complex dynamics online, and the dynamics of cultivating communicative interaction and response”. Published in 2015, Whitney Phillips’s This is why we can’t have nice things: Mapping the relationship between online trolling and mainstream culture is one of the first academic trolling studies to map the relationship between online trolling and mainstream culture. Phillips explains the evolution of the term troll from its origins of mythology in early periods, and fishing in 2003, to its use to describe vile cyber bullying in 2007 (p. 11). Phillips primarily finds “an increasing number of academics and journalists began joining the conversation, but initially there were very few sources to pull from” (p. 38) indicating that in 2015 supporting academic literature was still rare. Phillips operationally defines trolling for the purpose of her study, her definition of trolling as TBA was deemed too narrow to
capture the experiences of journalists in Australia so Johnston’s (2014) was used. While Phillips defines trolling, solutions to the issue do not fall within the scope of her study, and she “looks forward to further research that tackles these unanswered questions” (p. 168). This study responds to this call.

2.3.2 Evolution of use in the Australian news media

Events that first brought the word trolling into news headlines is Australia concerned the 2012 attempted suicide and 2014 suicide of Charlotte Dawson (ABC, 2014). Since then, the use of the term seemed to proliferate but no analysis of this process had been conducted, so this became one of the aims of this study. Use of the word ‘troll’ has increased in both media reporting and newsroom dialogue. In 2012 coverage of her, Dawson flagged several key issues including discrepancies in the understanding of how to deal with trolling and sparked calls by the Federal Coalition for uniform laws across Australia (Connelly & Keene, 2012), leveraging her fame with accounts of personal abuse to inject trolling discussions into workplace dialogues beyond newsrooms. Concurrently, the trolling of the then Prime Minister Julia Gillard in 2012 was also influential and led to the first Australian regulation of social media companies (Matheson, 2013; Morrissey & Yell, 2016). From there onwards, trolling attacks became more widely reported, and the term ‘troll’ propelled itself into the Australian media spotlight as a new buzzword (Thompson, 2019). However, it was not until two years later when Dawson in 2014 took her own life, allegedly because of continued trolling abuse, that trolling cemented itself as a social issue, not just simply a form of cyber abuse. To date, the use of the word ‘troll’ in the Australian news media has little academic literature. This study responds to this omission and begins to map the evolution of the term from the first mention of the word ‘troll’ in 2009, to 2015 through thematic content analysis described in the research design section 3.2.1.1 of this thesis (on p. 32).

2.3.3 Evolution of academic literature on trolling

Early discussions of trolling in academic literature emerged in digital communication studies in 1997 (Tepper, 1997; Donath, 1997) that recognised the key role that deception played in trolling. In further exploration of this idea,
Dahlberg (2001) said:

Deception in deliberation occurs in situations in which a participant intentionally misleads others into believing that intentions, needs, desires, and interests have been honestly presented. This is quite a widespread problem in online discussion groups. Sometimes called trolling, identity deception in cyberspace aims to embarrass, anger, and disrupt. It is often undertaken merely for amusement but is sometimes driven by more ‘serious’ motives including political goals.

(Para. 34)

When this study commenced in January 2015, research about trolling in journalism studies was still developing in peer-reviewed literature. Although media professionals agreed about its prominence and had heard of trolling anecdotally, no data mapped the emergence, impacts of and responses to trolling in the Australian news media, which this study aimed to do. The escalating frequency of trolling had heightened public awareness of the problem and, subsequently, began to attract academic attention to try to understand and resolve the phenomenon. Initial research progressed from trolling within online gaming communities (Thacker & Griffiths, 2012; Jane, 2015) to encompass all social media platforms (Alim, 2014; Tsantarliotis, Pitoura & Tsaparas, 2016). 2015 emerged as a year of notable scholarly contribution. Phillips (2015) summarised history and suggested the behaviour reflected a cultural problem. Also in 2015, research emerged from Australian journalist, award-winning author, and internationally renowned scholar of misogyny, gender, and technology-facilitated violence, Dr Emma Jane who reported, “a dramatic increase in a type of vitriolic discourse notable for its hostile affect, explicit language and stark misogyny” (Jane, 2015, para. 1). By 2016, her book, Misogyny online: A short and brutish history, further explored the worldwide phenomenon of gendered cyberhate as a significant discourse that had been overlooked and marginalised. In a major contribution, Jane's (2018) research identified online abuse as an occupational health and safety issue.

Criticism emerged in 2017 with Beckett (2017) who suggested “conflating jokes and death threats made it difficult to legislate or create effective polices” (Para. 8) and condemned the media’s misuse of the term trolling. By this time, research had
widely established that women were more likely than men to receive insults or threats (Binns, 2017; Jane, 2016) and Binns (2017) further found “women had stronger emotional reactions to abuse” (p. 183). Research progressed from documenting and defining the problem, to discussions about underlying root causes that were likely embedded far deeper within societal ideologies of feminism, racism and other forms of discrimination against minority groups (Jones, Trott & Wright, 2019; Schapals & Bruns, 2019). Claire Wolfe’s Online trolls, journalism and the freedom of speech: Are the bullies taking over? drew particular attention to the threat trolls pose to feminism and the freedom of speech. As research on comment moderation emerged as both a newsroom practise and a topic of research (Barnes, 2018; Gardiner, 2018), anti-feminist attitudes were again revealed. Research by Gardiner (2018) found that moderated articles written by females attracted a higher percentage of blocked comments regardless of the subject of the article. Labelled as “the paramedics of the internet” (White, 2022, para. 1), content moderators were a temporary coping strategy, rather than an effective deterrent, and continual exposure to hostile discourse began to have harmful impacts on the moderators (Ruckenstein & Turunen, 2019; Steiger et al, 2021). Although increased engagement with audiences in comment streams likely has economic benefits, Wright, Jackson, and Graham (2019) reported that trolling made online interactions more difficult for journalists, through time constraints and lower perceived value. Research by Masullo Chen and Chen (2019) also discussed the interplay between news overload and consumption. In what Gardiner (2018) described as a paradoxical challenge for comment space, Wright, Jackson and Graham (2019) concluded the costs for journalists were generally thought to outweigh the benefits of participation. In further support, research by Lewis, Zamith and Coddington (2020) found the more journalists were trolled, the more likely they were to perceive audience interaction as less valuable. Bossio and Holton (2021) found hostile interactions with audiences online has increased anxiety and frustration. As a result, journalists’ have changed the way in which they use social media by disconnecting from, rather than terminating, their accounts (Bossio & Holton, 2021).

More recent studies continue to frame trolling as a complex feminist issue (Duyn,
Peacock & Stroud, 2019) with Lewis, Zamith and Coddington (2020) reporting “as greater targets of the worst forms of abuse, women face a greater burden in deciding if and how to respond to online harassment” (p. 1047). Research has progressed from describing the problem and its effects on stakeholders, to potential strategies to reduce trolling online. Training and education emerged as a critical role to reduce the impacts of, and exposure to, trolling in journalism practise. Research by Martin and Murrell (2021) suggested dialogic interactions are a fundamental part of the future of journalism and, therefore, trolling needs to be ethically managed. While some journalists understand the relationship between resilience training and positive mental health, the majority are not clear about its potential and how it might be taught (Martin & Murrell, 2020). Journalism education needs to be innovative and transformative to address the normalisation of dialogic interaction on social media and the challenges it presents (Martin & Murrell, 2021). While research has not found a universal solution, the ‘wickedness’ of social problems means there may never be one (Camillus, 2008). Research is yet to define a systems-based approach to trolling that addresses key stakeholder groups and their interdependences. This study begins to explore potential sites for intervention to support a public-service profession where trolling has emerged as a widespread yet unresolved issue.

2.4 New media and trolling

2.4.1 News transition from print to online
News transition from print to online has been the focus of research since the 1990s. Although some researchers such as Bowd (1997) suggest “online newspapers are, in the age of computer communications, a logical successor to print” (p. 58), the presence of both in the current media landscape suggests there is some overlap in the progression from paper to online news. Bowd describes the strength of online news as its “ability to include vast amounts of news and information” (p. 60) and distribute this information efficiently without the time constraints and geographical restrictions of print news. For some time, commentators (Saffo, 1992; Matheson, 2004; Gillmor, 2004) welcomed the internet as a medium that promotes active participation rather than passive
consumption, and, as a result, has the potential to help create a more democratic and representative public sphere (Thurman & Hermida, 2010). The popularity of online news is supported by researchers Liu, Tseng and Chen (2015) who proposed news distributed via social media has become an “important communication platform in our daily life” (p. 1) with researchers Shamma, Kennedy and Churchill (2010) further claiming that “microblogging concurrently with live media events is becoming commonplace” (p. 331) and a habitually accessed news source. Although a fundamental role of journalism is to create public debate about significant societal issues, Martin and Murrell (2021) highlighted journalists have faced considerable challenges in generating productive, civil online conversations with their audiences. Bossio and Holton (2021) suggested the labour required to engage in social media practices leads to ‘social media fatigue’ (p. 2476) as journalists are required to navigate the incivility of dialogue. Some news organisations have been unable to sustain the cost of moderation, legal and reputation risks (Huang, 2016) associated with comment streams, that attract audience critique, aggressive commentary, information overload and security concerns (Martin & Murrell, 2021). Bossio and Bebawi (2016) reported most news organisations recognise the importance of social media to the production and dissemination of news, and therefore encourage active use by journalists to promote to and engage with their audiences. This participatory journalism creates “collaborative initiatives” (p. 150) that encourages citizens to contribute to stories that supplement professional journalistic work. Nora Martin (2015) reported digital journalism facilitates richer and more expansive storytelling, with connectivity between experts, journalists, and the public. In 2022, social media plays a pivotal role in how Australian journalists and news media organisations engage with their audiences. With the evolving business model of digital media, continued research is required.

2.4.2 Interactivity and trolling

In the online environment, interactivity immerses content creators and readers within a two-way communication model where readers can respond to a news article instantaneously in an interactive environment, predominantly designed to engage readers more than traditional media sources were able to. While more
recently in 2021 Martin and Murrell discuss dialogic journalism as a framework for understanding the dynamics of newsrooms interactions, this study uses interactivity to capture the broad shift from publications involving one way communication to the opening of feedback from audiences. Vine (2012) claims that newsrooms regard interactivity as a beneficial technique for news distribution, suggesting “editors appear to be enthusiastic about social networking capabilities, and are embracing, or ready to embrace, interactivity such as that offered by Facebook and Twitter” (p. 170), with most Australian news organisations embracing interactivity as an asset to news circulation, including The Sydney Morning Herald, Herald Sun, The Age, ABC, BBC and The Guardian. These evolving and proliferating interactive news forums analysed and categorised in greater detail by Martin and Murell (2020 & 2021) encouraged readers to use promotional features including ‘share’, ‘like’, ‘follow’ and ‘comment’ in response to articles about topics of personal interest in order to connect with other readers of similar interest, and thereby further increasing readership. Carey (2014) describes these two-way communication models as having “high levels of perceived satisfaction” (p. 121) and links this to higher news readership retention rates:

Many new companies have recognised the potential of harnessing social media interactivity to increase exposure to new stories and create online following . . . many researchers have placed interactivity among the key components necessary to understand new media technologies.

(Carey, 2014, p.121)

The interactive function that online news provides can be leveraged as a marketing strategy to promote an article to its target audience. Newsrooms monitor this interactivity to structure and refine future articles to those best received by the public, to ensure maximum readership exposure. The dynamic relationship that social media platforms permit provides readers with a sense of their importance, thus building rapport between the reader and the news provider and creating a sense of community between like-minded readers (Bossio & Holton, 2021; Bossio & Bebawi, 2016). However, trolling has emerged as a problem resulting from this online interactivity.
2.5.3 Anonymity and trolling

As a mechanism for egalitarian debate, anonymity can be incredibly valuable (Arvanitakis, 2015), yet as a facilitator of intense vitriol, anonymity threatens to undermine its central democratic function (Priestley, 2015). Anonymity provided by social media platforms has become a springboard for inflammatory comments, and when left unmonitored, comments have the potential to intensify, with extreme examples illustrating the ferocity and vulgarity trolling can escalate to: “holes like this make me want to commit rape out of anger” (West, 2015, para. 1) “choked to death with Satan’s dick” and “you are clearly retarded, I hope someone shoots then rapes you” (Ryan, 2014, p. 2 & p. 8). Trolling behaviour can be positioned along a continuum of harassment in online environments that also includes cyber bullying. Suler’s (2004) The online disinhibition effect highlighted that inappropriate vocabulary is more frequently and intensely used in online communication channels than in face-to-face ones, and further suggested those who use the online realm for communication “loosen up, feel less restrained and express themselves more openly” (p. 321), primarily because of the anonymity afforded by online platforms. Martin (2013) stated that unacceptable behaviour is further encouraged by removing the risk of “social sanctions” and “physical assault” (para. 3) when communicating online, encouraging the provocative vocabulary and vicious nature of trolling. Martin added that the absence of an immediate consequence further promotes trolling: “common wisdom dictates that people are more aggressive, rude and forthright online because they’re anonymous and can act as unpleasant as they like without immediate consequence” (para. 3). An inconsistency of consequences between websites, and the absence of any consequences at times, indicated as areas of concern and potential further exploration.

2.5 Trolling victims

2.5.1 Trolled journalists

While feminist writers such as Jezebel’s Lindy West (2015) flagged that dealing with trolling had become a “common part” (para. 2) of journalism practice, and a frequently overlooked workplace issue, its ongoing presence as a problem was
confirmed by Martin and Murrell (2020 & 2021) and Lewis, Zamith and Coddington (2020). One of the first places that trolling established itself as a notable workplace issue was in the newsroom of the feminist led publication Jezebel, in August 2014. Prior to this, only sporadic examples of trolling were published globally. Speaking out about the problem in an article letter, the feminist publication states:

This practice is profoundly upsetting to our commenters who have the misfortune of starting their day with some excessively violent images, to casual readers who drop by to skim Jezebel with their morning coffee only to see hardcore pornography at the bottom of a post about Michelle Obama, and especially to the staff, who are the only ones capable of removing the comments and are thus, by default, now required to view and interact with violent pornography and gore as part of our jobs. None of us are paid enough to deal with this on a daily basis.

(Jezebel 2015, para. 3)

By 2014, rape and death threats, such as those received at Jezebel, were becoming more common in online discourse (Ryan, 2014). Other prominent trolled journalists began to speak out including Finnish YLE journalist Jessikka Aro who was investigating the existence of pro-Russian troll factories in 2014 when she was targeted by trolls, both online and in real-life in, with an extensive harassment campaign that continued for years (Yates, 2017). By 2015, West (2015) suggested that being trolled had become a daily occurrence for practicing journalists and had developed into an industry norm. “I’m told, a constant barrage of abuse is just part of my job. Shrug. Nothing we can do. I’m asking for it, apparently” (para. 1). West’s sarcastic undertones illustrate her resentment of the onslaught of acrimonious online abuse when simply doing her job. Ryan (2014), a colleague at Jezebel, highlighted that West was not alone in the barrage she endured, saying, “like most women on the internet, I get told that I should be raped or murdered or set on fire... more frequently than most people would be comfortable with” (para. 2).  

Since West (2015) and Ryan’s (2014) articles were published, a number of reports have backed their claims that the first trolling victims, in overwhelming numbers, were journalists (Binns 2017; Gardiner, 2018; Lewis, Zamith & Coddington, 2020; Martin & Murrell, 2020; Martin & Murrell, 2021). With sarcasm, West further asserts that harassment of this nature would not be acceptable within any other
profession.

Being harassed on the internet is such a normal, common part of my life that I’m always surprised when other people find it surprising. You’re telling me you don’t have hundreds of men popping into your cubicle in the accounting department of your mid-sized, regional dry-goods distributor to inform you that – hmm – you’re too fat to rape, but perhaps they’ll saw you up with an electric knife? No? Just me?

(West, 2015, para. 2)

In addition, West argues trolling is unjustly accepted as a newsroom inconvenience, and this grave understatement does not accurately reflect the severity and scope of the problem. The issue is that newsrooms and media companies are not doing enough to document and respond to abuse (Martin & Murrell, 2020; Martin & Murrell, 2021) and the formation of policy responses is still underway. Some approaches have been tried but there is not yet a broadly recognised best-practice response. Media coverage of trolling debates has encouraged critical dialogue to spill into academic discourse, exacerbated by the fact that journalism is going through a period of unprecedented turmoil driven by technological change (Tanner et al, 2014). Trolling discourse is not sequential; it is messy, and the debate now seeps through professions beyond journalism and extends beyond local borders, now being evident on a global scale.

2.5.2 Trolling and gender

Although the growth of the digital landscape offers new opportunities for feminism, enabling those who were traditionally excluded from public communication practices, such as women, access to a platform encouraging democratic public civil engagements, it is also tainted by familiar forms of misogynistic abuse and harassment. Notable Australian trolled journalist Clementine Ford (2017) said the backlash that inevitably accompanies online discourse finds new and creative ways to manifest (para. 2). In 2016, the Media Entertainment Arts Alliance survey found 41% of women said they had been harassed, bullied or trolled on social media, while engaging with audiences; several were silenced, or changed career. As females are more frequently victims of targeted abuse, particularly those who express feminist views (Ford, 2017;
Noyes, 2017; West, 2015; Ryan, 2014; Aune & Refern, 2013; Penny, 2012; Tyler, 2012) increased academic, media and political attention echoes longstanding gender inequalities: “Online abuse of women may be a twenty-first century issue, but its roots grow from much older forms of sex-based harassment and misogyny” (Tyler, 2012, para. 1). Tyler also suggests misunderstanding the problem, independent of its patriarchal origins, “overlooks and trivialises the harm it causes” (para. 1). In support, Jane and Vincent (2017) argue that misogyny online constitutes attempts to silence women through intimidation, coercion and self-censorship (p. 68). Consensus from Lewis (2018), Jane and Vincent (2017) and Priestley (2015) is that trolling is not simply a new phenomenon, but a new face of more enduring patriarchal practices, including the ignominious history of the silencing of women. Priestley further suggests women who express their right to have a voice online, “threaten the longstanding, traditional notions of power” (para. 1). Trolling has resurfaced feminist discussions in popular culture (Redfern & Aune, 2013); however, Lewis (2018) warns that while the returned attention to feminist concerns is welcome, experience of patriarchal systems teaches us to be cautious about hopes of complete elimination.

Voices from early victim accounts echoed in global news articles, stressing the psychological impact on victims. The pioneers of public awareness and debate about trolling in the Australian news media were women. Four years before Lindy West’s (2015) article, which prompted this research, was published, Australian trolled journalist Karalee Evans (2011) voiced the same concerns, asking “when did faceless men decide it was acceptable to take it upon themselves and threaten women online with death threats, rape threats, violence and sexism?” (para. 5). Evans (2011) suggests trolling is an extension of sexism, adding online harassment “is not a new thing, it existed even in the days of MySpace and GeoCities, what seems to be on the rise is compliance trolling and the phenomena of anonymous digital misogyny” (para. 5). In support, other Australian trolled journalists began to respond, with Elmo Keep (2012) speaking out:

What many women writers have in recent days come forward to share as the online abuse they suffer at the fingertips of losers, is not funny. Incitements to rape are not funny. To threaten someone with any kind of physical
violence is not funny. To wish someone, or their family or pets dead, is not funny. To invite someone to commit suicide is not funny. It’s not funny no matter what the context, or the person saying it. It’s just not, it’s missing that essential humour ingredient: being in any way funny (Keep, 2012, para. 6).

Australian trolled journalist Meagan Tyler (2012) also spoke out, saying “trolling has captured the attention of the mainstream media” and the “cyber abuse of women is finally getting noticed” (para. 2). By 2012, intensive misogynistic comments became the catalyst for the first Australian female journalists (Evans, 2011; Keep, 2012; Tyler, 2012) to speak out. From this came traction within the Australian news media and the debate snowballed (Thompson, 2019). The impact of trolling spread beyond the newsroom and into the victim’s personal lives, inflicting anger and hurt. Australian trolled journalist, Ginger Gorman, voiced further concerns about the professional and economic harms of trolling to women, stating “gendered cyberhate silences women and constrains their ability to find jobs, market themselves, network, socialise, engage politically and partake freely in self-expression and self-representation” (Gorman, 2019, p. 71). Trolling has cemented itself as a significant current workplace issue for Australian journalists in 2022 that requires academic research to inform each stakeholder group of the best response strategies. Although Evans (2011) called for answers to reverse this vile and abusive digital sexism (para. 18) back in 2011, academic research has not yet provided solutions to effectively protect Australian journalists from trolls. This study begins to address this continued call. When this study commenced in 2015, the impetus behind it was the calls from journalists, combined with emerging data from the Media Entertainment Arts Alliance (2016) which indicated it was a complex problem. Ackoff’s framing of social problems posited that other positions of power in the field were worthy of analysis, and so the tiered approach was adopted. Since the commencement of the study, the body of research has grown and more recent studies (Binns 2017; Bossio & Holton, 2021; Bossio & Debawi, 2016; Gardiner, 2018; Lewis, Zamith & Coddington, 2020; Martin & Murrell, 2020; Martin & Murrell, 2021) have found more evidence of the need for greater understanding and action.
Chapter 3
Methodology

This chapter introduces the specific research questions and the methodological approach used to address them. As the literature review demonstrated, by the commencement of this research in 2015 trolling in the news industry was being described as a social problem and attempts to understand it were in nascent stages. This study sought to contribute to understanding trolling as a social problem. The chapter will also discuss ethical considerations and the limitations of this research process.

3.1 Over-arching research question
How did trolling emerge in the Australian news industry, what were the impacts on journalists, and stakeholders' responses to it?

Ackoff’s system’s approach to societal problems (1974) was adopted to ensure the investigation was holistic and did not miss key stakeholders or focus inappropriately on specific stakeholders such as victims. Adopting this model generated a set of research questions focussed on the different levels in Ackoff’s model. In this way the model is a true framework for the thesis. The questions derived from consideration of stakeholders at each level of the model are listed below.

3.1.1 Research sub-questions

Emergence - Question A
How can the evolution of trolling in the Australian news media between 2009 and 2017 be described?

Impacts - Question B
What are the impacts for trolled journalists in the Australian news media?
Campaign responses - Question C

How can the evolution of anti-trolling campaigns within the Australian news media be described?

Twitter responses - Question D

What policy and feature changes were implemented by Twitter in response to trolling from its launch in 2006 to 2019?

Facebook responses - Question E

What supportive strategies, operational modifications and corporate reform were implemented by Facebook in response to trolling?

Newsroom responses - Question F

What strategies were implemented by Australian news organisations in response to trolling?

Legal responses – Question G

What Australian laws can be used by journalists to prosecute a troll?

Each of these questions is addressed in a chapter of the thesis. The chapter order is grounded in the data from chapter 4.1 which focussed on the emergence of trolling and was the first part of the research completed. From this data it became clear that victims were initiating calls for the issue to be addressed and were a logical starting point in the stakeholder analysis. Their views are discussed in chapter 4.2. The emergence data showed that their calls coalesced into campaigns which are the focus of chapter 4.3, and pressured social media companies to respond, with Twitter the focus of chapter 4.4 and Facebook the focus of chapter 4.5. The emergence data also showed newsrooms as explored in chapter 4.6 were slower to release information about anti-trolling strategies and their response times were slower as strategies were backed by academic research which takes time to formulate. The emergence data showed legal responses in chapter 4.7 emerged last, with the nationally agreed upon Model Defamation Amendment Provisions 2020, Online Safety Act 2021, and Social Media (Anti-Trolling) Bill 2022.
that began to more adequately address trolling.

3.2 Research design

3.2.1 Methodological approach

Although predominantly using qualitative data, proposed as a viable means of research for media studies in an array of academic literature (Altheide & Schneider, 1996; Brennen, 2017; Crotty, 1998; Gough-Yates, 2003; Lindoff, 1995; Percy, Kostere & Kostere, 2015), this multi-method study uses the data collection methods of in-depth interviews, archival evidence, and thematic content analysis, each of which incorporates a quantitative and qualitative data component, discussed later in further detail.

Weerakkody (2015) suggests a mixed-method approach increases the validity and reliability of data collected, and enables a more comprehensive discussion, understanding and critique of research (p. 32), known as methodological triangulation (p. 32) or crystallisation (Wimmer & Dominick, 2011, p. 64). To provide a comprehensive analysis, triangulation is adopted to incorporate all three data collection elements of in-depth interviews, archival evidence, and thematic content analysis. The complexity of trolling invites meticulous investigation, which can only be achieved using different and complementary methods to ensure each aspect of the multifaceted problem is addressed.

First introduced by sociologists Glaser and Strauss (1967), and further developed by Strauss and Corbin (1990), the grounded theory approach to the analysis of qualitative data, selected for this study, refers to the method of using a systematic set of procedures to code data into named categories to discover patterns among them (Weerakkody, 2015, p. 301). Identifying correlations allows social patterns to be conceptualised and compared. An inductive approach is first used to form substantive codes for initial data, and from this emerging theory, the next step of the data collection can be deduced. Glaser and Strauss’s (1967) work *The discovery of grounded theory* is one of the most cited methodologies in social sciences, cited more than 149,447 times at the conclusion of 2021 on Google scholar by the
conclusion of this study. Grounded theory is used as a methodology in journalism research, including in the widely referenced PhD theses of Martin (2004) and Glynnild (2006).

Google was emerging as a useful research tool in academic literature (Piasecki, Waligora & Dranseika, 2018; Robertson et al, 2018; Ormen, 2015). Although personalisation algorithms may change which results are presented as top, it would not affect the total number of articles retrieved by the Google search. Bias risk was limited as no Google accounts were active at the time, and the research was conducted before 2016 when Google targeting was less sophisticated.

3.2.1.1 Emergence – Question A

The research design strategy which addresses: *How can the evolution of trolling in the Australian news media between 2009 and 2017 be described?* applies the thematic content analysis method. To map the evolution of trolling in the Australian news media, thematic content analysis, a descriptive collation, analysis and presentation of qualitative data (Anderson, 2007), was selected to analyse three prominent Australian news websites with different owners from 2009 to 2017: smh.com.au; news.com.au; and abc.net.au. Using Google as a search tool, the search inquiry `site:smh.com.au "troll" -"movie" -"fishing"` was applied for each of the news websites. Results from this search provided a list of trolling-related news articles for each calendar month between January 2009 and December 2017 inclusive. The monthly frequency of trolling-related articles was recorded for each news website, eliminating any articles unrelated to the research topic, providing quantitative raw data. The top three Google ranked trolling-related articles for each news website were identified and recorded in Microsoft Word, providing qualitative raw data. Applying thematic content analysis, suggested by Gering (2015) as an appropriate method for arranging texts into a unified and generalised coding structure which, in turn, can be transformed into a database, the 528 articles were coded onto an Excel spreadsheet, using pattern code or open coding (Miles & Huberman, 1984, 1994; Weerakkody, 2009). The raw data set was cleaned according to coding principles for coding rigor before analysis and discussion.
3.2.1.2 Impacts – Question B

The research design strategy which addresses: *What are the impacts for trolled journalists in the Australian news media?* applies the mixed methods of thematic content analysis and in-depth interviews. A deliberate network sample was made of ten industry professionals, between the ages of 23 and 65, five males and five females. Each had at least five years’ experience in the journalism field and self-selected as being a trolling victim. Three participants had written publicly about being trolled, two were known to supervisor Dr Kayt Davies, three were identified by contacting the Chief of Staff from *ABC, 9 News Perth, Seven West Media,* and *Network Ten* in Subiaco, and two were known to participants via snowball sampling. These participants were selected for semi-structured interviews of approximately an hour duration. The participants were asked questions about their personal experiences with trolling in the workplace. Participants reflected on any negative personal impacts, identified any workplace support systems in place, and specified any training they received from university or their employers. Responses were then analysed for emerging themes in conjunction with the theoretical frameworks employed in this thesis.

3.2.1.3 Responses to trolling

Through a grounded theory approach described by Weerakkody (2015), the study “develops theories to explain a specific real-life problem or phenomena (trolling) by systematically analysing patterns, themes and common categories observed in qualitative data” (p. 301). From the systematic data analysis derived from the thematic content analysis and in-depth interviews of *Questions A* and *Question B,* six key stakeholders groups emerged as actively responding to trolling within Australian news media being: victims (addressed in *Question B*); anti-trolling campaigns; social media platforms; newsrooms; the law; and educational resources. These five remaining groups will be addressed individually as part of this research’s use of Ackoff’s (1974) systems approach, within five subsequent research designs, each exploring a different level of the social problem trolling.
3.2.1.3.1 Campaign responses – Question C
The research design strategy uses mixed methods of thematic content analysis and archival evidence. Using Google as a search tool, the terms: “troll” + “campaign” were investigated with the location set Australia. From analysis of the first ten pages, totalling 100 articles, 13 significant campaigns were identified as archival evidence. Activist responses were plotted on a timeline that documents and illustrates the evolution of anti-trolling campaigns in the Australian news media. Campaigns drew attention to risks of allowing trolling to go unchecked, provided context for newsrooms experiences, and illustrated that the problem was systemic, rather than discrete episodes, which was best addressed through the application of a systems approach.

3.2.1.3.2 Twitter responses – Question D
The research design strategy which addresses Question D: What policy and feature changes were implemented by Twitter in response to trolling from its launch in 2006 to 2019? applies mixed methods of thematic content analysis and archival evidence. The study used collected data from chapter 4.1 (on p. 39) as a starting point. Of the 528 articles in the thematic content analysis, 17 mentioned Twitter response to trolling on its platform in some capacity and were deemed relevant for inclusion. Of these, five further discussed or made comment on aspects of particular response strategies. To obtain further data, Twitter’s website was explored, and relevant information was retrieved from three sections: About Us, Help Centre and Blog. On each of these pages the search function was used to investigate the terms: “troll” “bully” and “abuse”. This led to a further 16 relevant articles which were included for analysis. Although Google was also used as a search tool with the search inquiry: “troll” + “Twitter” + “response” and retrieved 3,120,000 results, on closer analysis of the first ten pages totalling 100 articles, only one additional response strategy was found.

3.2.1.3.3 Facebook responses – Question E
The research design strategy which addresses: What supportive strategies, operational modifications and corporate reform were implemented by Facebook in response to trolling? applies mixed methods of thematic content analysis and
archival evidence. Of the 528 articles in the thematic content analysis of this study, 31 articles mentioned Facebook’s response to trolling on its platform in some capacity and were deemed relevant for inclusion. Of these, 14 articles discussed or made commented on aspects of particular response strategies. To obtain further data, Facebook’s website was explored, and relevant information was retrieved from three sections: About Us, Safety Centre, and Help Centre. On each of these pages the search function was used to investigate the terms: “troll” “bully” and “abuse”. This led to a further 25 relevant articles which were included for analysis. Although Google was also used as a search tool with the search inquiry: “troll” + “Facebook” + “response” and retrieved 3,520,000 results, on closer analysis of the first ten pages, totalling 100 articles, no new response strategies were found that had not already been identified in the previous methods of this chapter.

3.2.1.3.4 Newsroom responses – Question F

The research design strategy which addresses: What strategies were implemented by Australian news organisations in responses to trolling? applies mixed methods of thematic content analysis and interviews. Four industry experts were selected and contacted, two were available to interview. All selected industry experts were Australian. Media academic Dr Jenna Price, who in 2019 completed a PhD on feminist digital activism in Australia and its account of fatal violence against women, and lecturer and media academic whose research focused on online community management, Dr Jennifer Beckett, were both interviewed. In addition, academic and journalism educator Julie Posetti who co-wrote the handbook for online safety in journalism practice (Ireton & Posetti, 2018), and UNESCO researcher and digital journalism academic Associate Professor Fiona Martin, were unavailable to interview. Participants were reached through academic and industry contacts. The semi-structured interviews went for approximately one hour and provided both quantitative and qualitative data. Each participant was asked to respond with an agreement level represented by a numerical value on a scale of one to ten, quantitative data regarding Australian newsroom guidelines was provided. Predominantly employing open-ended questions, robust qualitative data was acquired. Newsrooms ABC, 9 News Perth, Seven West Media,
and *Network Ten* in Subiaco were contacted and asked about their policies in response to trolling. *ABC* policies were found on their website. *Seven West Media* and *Network Ten* dealt with matter internally on a case-by-case basis, and *9 Perth* (formerly *Fairfax Media*) refused twice to provide a copy of their handbook for online safety produced by Julie Posetti. In addition, developing patterns regarding the use and effectiveness of newsroom support systems were drawn upon from the in-depth interviews of *Question B* and the thematic content analysis for comparison and discussion.

### 3.2.1.3.5 Legal responses – *Question G*

The research design strategy which addresses: *What Australian laws can be used by journalists to prosecute a troll?* applies mixed methods of thematic content analysis and interviews. In interviews with two selected Australian media law experts lawyer Mark Polden and Professor Mark Pearson, authors of the most widely used media law textbook in Australia *A Journalist’s Guide to Media Law: A Handbook for Communicators in a Digital Age* (2019). Participants were asked to consider which current Australian laws could be drawn upon to prosecute trolls. Participants reflected on any notable prosecution examples, discussed legal pathways for trolled victims, and considered the legal impacts of trolling in Australian newsrooms. In addition, participants reflected on the evolution of these laws from 2009 to 2019, and concluded with personal judgements concerning the law’s accessibility, effectiveness and adequacy in a society where trolling remains pervasive. Through investigation of legislative websites, the collection of archival evidence revealed which Australian laws could be drawn upon to prosecute a troll.

In the subcategories of criminal and civil, federal, and state, the number of Australian laws were tallied to generate quantitative data. A description of these laws provided qualitative data for comparison and discussion. From this data set, one criminal law, and one civil law flow chart were constructed to provide a practical reference for a trolled journalist to follow, while highlighting notable areas of concern within the Australian legal system.
3.3 Ethical considerations

An ethics application was obtained from Edith Cowan University after review by the university’s Human Research Ethics Committee. Research did not commence until ethics approval was granted. There was considered to be minimal risk to all participants.

This study took into consideration the sensitivity of responses and implemented full anonymity for all victim interviewees. This allowed unaltered descriptions of trolling recounts, without fear of backlash from trolls, judgement by peers, and importantly, repercussions from their employers. Results were deidentified through coding referred to participants by number within the text of the research ensured the best platform for data accuracy and depth of rich data.

3.4 Limitations

The findings are limited to the Australian news media. A comprehensive investigation about the extent of trolling globally fell beyond the scope of this research. While Suler (2004) and Barnes (2018) have sought to understand the motives and behaviour of trolls as a stakeholder group, it was deemed beyond the scope of the study to include analysis of them. The research period of 2015 to 2021 is also a limitation of the study, and further investigation about the application and effectiveness of Australia’s emerging trolling laws in 2022 (as introduced and discussed in chapter 4.7 on p. 265) is recognised as a valuable area of future analysis. As qualitative responses reached a point of saturation, ten victim impact interviewees in chapter 4.2 were considered adequate to frame the personal and professional implications of trolling in contemporary journalism. However, greater depth could be achieved through supplementary academic interviews that further discuss the implementation and effectiveness of trolling response strategies by key stakeholder groups. In the newsroom chapter 4.5 two of the industry experts selected to interview were not available, time limitations restricted more extensive interviews, and newsrooms refused access to policies. While grounded theory could encompass the study of everything, the content analysis research process required selections to be made to control the scope of the undertaking. These selections and limitations are described in their relevant
chapters. I hope this study will indicate directions for future research to address the escalating need for a synthesised approach to trolling in both Australia news media, and newsrooms around the world.
Chapter 4.1
Emergence

Results and Discussion

A content analysis of 528 news articles

Question A - How can the evolution of trolling in the Australian news media between 2009 and 2017 be described?

Abuse and harassment have become commonplace for many Australian journalists, not only because of the interactive function of Web 2.0, but as a broader reflection of current politics, economics, and ideologies. Contemporary journalists are increasingly obligated to leverage the interactive internet functions for community building, self-branding, and readership engagement to increase circulation that trolling itself derives from (Bowd, 2016). In an epidemic of harassment, the study reveals some of the first documented insights, trends and themes emerging and further intensifying within the frontline battle faced by journalists as a collective part of modern journalism practice in Australia.

To map the evolution of trolling within the Australian news media, thematic content analysis, a descriptive collation, analysis, and presentation of qualitative data (Anderson, 2007) was selected to analyse articles about trolling from three prominent Australian news websites with different owners from 2009 to 2017: smh.com.au; news.com.au; and abc.net.au. Using Google as a search tool, the search inquiry site:smh.com.au "troll"-"movie"-"fishing" was applied to each of the news websites. Analysis of more than three news websites fell beyond the scope of this study. Results from this search provided a list of trolling-related news articles. These articles were commonly journalists self-reflecting about their trolling experiences, often from a feminist perspective. The frequency of articles for each calendar month between January 2009 and December 2017 inclusively were recorded for each news website, eliminating any articles unrelated to the
research topic, providing quantitative raw data. The top three trolling-related articles as ranked by Google search for each news website were identified and recorded in Microsoft Word, providing qualitative raw data. Applying thematic content analysis, suggested by Gering (2015) as an appropriate method for arranging texts into a unified and generalised coding structure which, in turn, can be transformed into a database, the 528 articles were coded onto an Excel spreadsheet, using pattern code or open coding (Miles & Huberman, 1984, 1994; Weerakkody, 2009). The raw data set was cleaned according to coding principles for coding rigor before analysis and discussion.

4.1.1 Quantitative data results
To find patterns in media coverage, a histogram was created using Microsoft Excel, plotting the frequency of trolling-related articles in the Australian news media between 2009 and 2017. On the Y axis, the total number of news articles containing the word ‘troll’ as listed by the Google search engine were recorded for each calendar month, and further cleaned to remove any irrelevant articles from the data set such as ‘fish trolling’. The data was displayed in monthly increments along the X axis to allow for comparison between frequency spikes and notable trolling incident dates. News.com is displayed orange, SMH in blue, ABC in green.

Figure 3
*Frequency of Trolling Articles in the Australian News Media*
Note: The histogram displays the total number of trolling-related news articles retrieved on the Google search engine for each month between 2009 and 2017. Each graph is displayed in parallel for instant comparison. ABC, shown in green, had the lowest frequency and was therefore displayed in the foreground of the graph. SMH, shown in blue, had the second highest frequency and was therefore placed in the middle ground of the graph. News.com, in orange, had the highest frequency and was therefore displayed in the background of the graph.

A collection of 528 articles was selected for content analysis. Before reporting the results of that study, data describing the set of articles was analysed. The line graph below shows a comparison between the frequencies of SMH, News.com and ABC trolling-related news articles in the thematic content analysis of this study. Displayed in yearly increments on the X axis, a linear regression line was plotted for each data set, indicating a positive correlation for each respective news website, meaning the number of trolling articles increased across all three news websites between 2009 and 2017. SMH is displayed in blue, News.com in orange, and ABC in green.

**Figure 4**

*Frequency of Trolling Related News Articles in the Thematic Content Analysis of this Study*
Chapter 4.1 Emergence results and discussion

Note: The data does not present the total number of trolling-related news articles in the Australian news media. The line graph displays the frequency of trolling-related news articles identified in the thematic content analysis of this study each year between 2009 and 2017.

Trendlines from the above graph show the data set is not ‘normalised’ and, therefore, parametric tests would not generate any valid findings. Mumby (2002) suggests parametric tests are considered by some academics to have more statistical power than nonparametric tests. But Kitchen (2009) argues that “nonparametric tests are much more powerful than their parametric counterparts when the underlying distributions are heavy tailed or have extreme skewness” (p. 571) and this skewness is evident in the data distribution of this study. Nonparametric tests that analyse median values, rather than mean values, are therefore better suited for this dataset. Bryman and Cramer (2009) and Panik (2012) consider nonparametric tests coupled with further qualitative analysis suitable for social science studies such as this doctoral study.

The content of the identified trolling articles in the thematic content analysis was further analysed to unpack details about the victims and the trolling comments received. The gender of the trolling victims identified in the thematic content analysis was analysed using the Mann-Whitney nonparametric test. The test was conducted. However, the male trolling victims identified in the thematic content analysis of this study scored 29.5 and the female trolling victims identified in the thematic content analysis of this study scored 51.5 respectively, meaning the gender distributions difference was not statistically significant (Mann–Whitney U = 17, n₁ = n₂ = 9, P < 0.05 two-tailed).
Figure 5

Gender Comparison of Trolling Victims in the Thematic Content Analysis of this Study.

Note: The data does not present the total number of trolling-related news articles on the Google search engine. The pie graph displays a gender comparison of the trolling-related news articles identified in the thematic content analysis of this study each year between 2009 and 2017.

To illustrate victim gender frequencies, a pie chart, displayed in Figure 5, above was created using Microsoft Excel. The pie chart above shows the ratio of trolling victims in the thematic content analysis of this study to be 46% female, 32% male, and 22% with their identity redacted. In the data-coding process, fourteen articles contained both male and female victims. These fourteen articles were subsequently included in the victim count for both genders, and one article contained a transgender person, and was included in the count for the gender they identified as, in this case female. Genders of 22% of trolling victims mentioned in the 528 articles are unknown, as they were not identified in original published article, or had redacted their identification from the article by anonymising the by-line after publication at the time of the thematic content analysis in 2017.
The line graph below shows the frequency of trolling-related articles containing rape threats in yearly increments between 2009 and 2017 and separated by gender. Male victim gender is moderately correlated with rape threats (displayed in blue), female victim gender is strongly correlated with rape threats (displayed in pink), and 100% of the 118 redacted identities (displayed in grey) were subjected to rape threats, indicating the strongest positive correlation.

**Figure 6**

*Comparison of victim gender and frequency of rape threats in the thematic content analysis of this study*

Note: The data does not present the total number of trolling-related news articles in the Australian news media. The line graph displays gender comparison of the trolling-related news articles identified in the thematic content analysis of this study each year between 2009 and 2017. Regression lines are not displayed beyond 0 on the Y axis as a negative number of identified articles is not a possibility.
The line graph on the following page shows the frequency of trolling-related articles of death threats over the same time period. Male and female positive correlation pattern trends are reflected in both rape and death threat graphs. However, the unknown gender trend shows a low positive correlation, with less reported incidences.

**Figure 7**

*Comparison of victim gender and frequency of death threats in the thematic content analysis of this study*

Note: The data does not present the total number of trolling-related news articles in the Australian news media, however, displays the gender comparison of the trolling related news articles identified in the thematic content analysis of this study each year between 2009 and 2017. Regression lines are not displayed beyond 0 on the Y axis as a negative number of identified articles is not a possibility.
The study considered classifying some types of trolling as more harmful or severe than other types as unethical, because all trolling is potentially harmful and can have differing levels of impact depending on the receiver’s personality and life experiences, their level of support, and the frequency of attacks. Therefore, the data set was coded into two categories of ‘insults’ and ‘threats of physical attack’. Trolling content considered sexist, misogynist, body shaming, homophobic, ableist, political or racist were categorised as ‘insults’, and trolling content containing rape and death threats were categorised as ‘threats of physical attack’. The study does not suggest that one category is worse than the other but has divided trolling articles into the two respective groups for the purpose of describing a pattern in the data set.

Excluding victims with redacted identities, binary logistic regression was used to analyse the remaining 425 victims. The results indicate that female victims are significantly more likely to receive insults than male victims: Female (χ²(1) = 1.246, p < 0.001) Male (χ²(1) = -0.124, p = 0.177). The results also indicate that female victims are significantly more likely to receive threats of physical attack than male victims: Female (χ²(1) = 1.246, p < 0.001) Male (χ²(1) = -0.124, p = 0.729).

However, with 22% of victims with redacted gender identities (totalling 118 people), the above results may not truly reflect the scope of the gender imbalance in being subjected to different types of trolling. Although results regarding this portion of data are not definitive, trend patterns can be reported for further analysis, comparison, and discussion about the extent of the problem. Of the 22% portion of redacted gender identities, Pearson’s correlation coefficient indicated a positive correlation between the number of trolling articles containing:

- rape and death threats, r = 0.832, n = 528, p < 0.001
- rape and sexist comments, r = 0.489, n = 528, p < 0.001
- rape and misogynist/anti-feminist comments, r = 0.337, n = 528, p < 0.001
- rape and body shaming comments, r = 0.252, n = 528, p < 0.001
- rape and homophobic comments, r = 0.172, n = 528, p < 0.001

This data suggests a significant portion, and possibly all, of the redacted gender
identities are likely to be female, further shifting the gender skew.

Figure 8

*A significant portion of the 118 victims with redacted gender identities are likely to be female*

<table>
<thead>
<tr>
<th>Correlations of redacted gender identities</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rape threats</strong></td>
</tr>
<tr>
<td>Death threats</td>
</tr>
<tr>
<td>Sexist comments</td>
</tr>
<tr>
<td>Misogynist/anti-feminist comments</td>
</tr>
<tr>
<td>Body shaming comments</td>
</tr>
<tr>
<td>Racist comments</td>
</tr>
<tr>
<td>Homophobic comments</td>
</tr>
</tbody>
</table>

Note: The positive correlations between articles containing rape threats in addition to death threats, sexist comments, misogynist/anti-feminist comments, body shaming or homophobic comments are highlighted in red.

4.1.2 Quantitative thematic analysis

**Frequency**

The histogram in Figure 3 (on p. 40) provides a visual representation of the emergence of trolling in the Australian news media. The first notable spike in September 2012 followed the attempted suicide of Charlotte Dawson on August 30, 2012. From then onwards, the term ‘troll’ snowballed until it was featuring frequently in mainstream media dialogue by the conclusion of 2017. The histogram also describes visual discrepancies between the three Australian news websites of the thematic content analysis with News.com publishing the most trolling related articles across the time-period totalling 839 articles, followed by the SMH with 644 articles, and notably far less frequently was the ABC totalling 250 articles. The thematic content analysis revealed News.com and the SMH often published trolling related articles quicker than the ABC due to an eagerness to publish news stories first. The importance of timeliness as a news value is evident throughout the news reporting techniques of the SMH and News.com, with successive articles often accumulating on the same trolling topic as additional
updates transpired. The ABC habitually provides a rich descriptive reporting technique, publishing articles less quickly but with more detail, including multiple comments and content from sources (Jolly, 2014).

Consistencies in the data shape of the three news websites suggest the number of trolling topics may be similar, however the number of corresponding articles to each trolling topic may vary due to differences in news reporting styles. The linear regression trendlines in Figure 4 (on p. 41) reflects a positive relationship, indicating the existence of the dependent variable: the frequency of trolling related articles, and the independent variable: time. It is important to highlight that regression analysis does not suggest a cause-and-effect relationship between the variables but does indicate the extent to which variables are associated with each other.

**Gender**

The pie chart in Figure 5 (on p. 43) provides a visual representation of the trolling victim gender imbalance, with 46% of trolling victims in the thematic content analysis of this study being female, and 32% male. From a feminist perspective, these numbers alone are alarming and evidence the misogynistic online culture described by West (2015), Ford (2017), Noyes (2017), Ryan (2014), Keep (2012) and Evans (2011). However, by unpacking the data, a potentially greater imbalance is revealed. 22% of the trolling related articles had redacted the victim’s identity in 2017 when the thematic content analysis data was gathered, and consequently exact gender analysis was not possible. These articles were commonly journalists self-reflecting about their trolling experiences, often from a feminist perspective.

By analysing trolling content in the 528 online articles, the study can begin to compare the potential victim impact of trolling experienced by people of different genders. The data indicates females are more likely to receive trolling rape threats, with 52 females and 13 males receiving comments of this nature. The data reveals 100% of the 118 gender redacted victims were subjected to rape threats, inferring a significant portion of this gender population may possibly be female, further
skewing the gender imbalance. The line graph in Figure 6 (on p. 44) additionally supports this interpretation, with the female line graph reaching a climax in 2016 and then falling in 2017, yet the male and redacted genders trolling frequencies both continue to rise. This abnormal decline in the number of recorded female victims receiving rape threats may indicate the time when female victims began to redact their identity due to the overwhelming impact of these comments, therefore subtracting from the female overall incidence rate and distorting the true gender imbalance. The line graph in Figure 7 (on p. 45) shows female victims are more likely to receive death threats in comparison to male victims. This supports the inference that a significant portion of the redacted gender population may be female, because the same abnormalities of data preceding 2016 is evident where the female trend again shows an unexpected decline when both the male and redacted gender victims continue to rise.

In summary, the binary logistic regression results suggest trolling content categorised by this study as either ‘insults’ or ‘threats of physical attack’ are more likely to be received by female victims. Furthermore, when the three coalescing dynamics are considered:

- the female victim linear regression trendlines in Figure 6 (on p. 44) and Figure 7 (on p. 45)
- thematic content analysis data showing 100% of the 118 redacted victim gender population were subjected to rape threats, and
- Pearson's correlation coefficient table in Figure 8 (on p. 47) indicating a positive correlation between rape threats and death threats, rape threats and sexist comments, rape threats and feminist comments, rape threats and body shaming, and rape threats and homophobic comments.

The data infers a significant portion of the redacted gender population may be female. Although inferences regarding this redacted gender portion of the data are not definitive, it is important to consider that if a significant portion of the redacted gender population is female as this study proposes, initial findings may not reflect the true scope of the gender imbalance.
4.1.3 Qualitative data results – a timeline of events

From the thematic content analysis of this study, a timeline of key trolling events in the Australian news media emerged, with phases of development identified as follows:

**Embryonic period (1982 to 1999)**

From the introduction of the internet in 1982 to the introduction of Web 2.0 in 1999.

**Incubation period (1999 to 2006)**

From the introduction of Web 2.0 to the introduction of Twitter in 2006.

**Manifestation period (2006 to November 2011)**

From the introduction of Twitter in 2006 to the viral anti-trolling campaign #mencallmethings by Sady Doyle on Twitter in November 2011.

**Social problem period (November 2011 to February 2014)**

From the viral anti-trolling campaign #mencallmethings by Sady Doyle on Twitter in November 2011 to the suicide of Charlotte Dawson in February 2014.

**Social epidemic period (February 2014 to June 2016)**

From the suicide of Charlotte Dawson in February 2014 to the election campaign of Donald Trump in June 2016.

**Saturation period (June 2016 to December 2017)**

From the election campaign of Donald Trump in June 2016 to the conclusion of the thematic content analysis in December 2017.
Figure 9

Chronological timeline of the evolution of trolling in the Australian news media derived from chapter 4.1 data.
4.1.4 Qualitative thematic analysis

In order to describe the evolution of trolling in the Australia news media, a timeline was constructed. Key events were identified to map the shape of the emergence of trolling, from which time periods were created for further analysis and discussion.

**Embryonic period (1982 to 1999)**

The *embryonic period* is identified as the time period prior to the introduction of Web 2.0 in 1999. In this period, the interactive internet was in its embryonic state and the key interactive functions of Web 2.0 required for trolling of journalists to emerge were limited. *ABC Online* launched its first website in 1995; however, it was not until 1997 that its first online discussion forum was introduced, followed by the *New York Times* in 1996, and *BBC Online* in 1997. The rise of the bulletin board system and forums emerged as a mechanism for interactivity and critique of journalists. News transition from print to online has been the focus of research since the 1990s, and although online news has been considered the “logical successor” to print (Bowd, 1997, p. 58), the presence of both formats in the current media landscape suggests there is some overlap in the progression from paper to online news. The ability to include vast amounts of news and information and distribute this information efficiently without the time constraints and geographical restrictions print news entails, are notable strengths of the online domain.

**Incubation period (1999 to 2006)**

The *incubation period* is identified as the time period from the introduction of Web 2.0 in 1999 to the launch of Twitter in 2006, which includes the launch of Facebook in 2004. This period saw an explosion of forums and later commenting sections as news media moved to a blogging style template for interaction. At this point, the technology was struggling with the interactive functions of Web 2.0, but interactive capacity was limited by the underdevelopment of social media platforms in existence. It emerged that guidelines were needed to keep online interactions with journalists civil during the time-period that Margo Kingston edited *WebDiary* for *The Sydney Morning Herald* in July 2000 (WebDiary, 2020).
Potential perpetrators, therefore, had the technological advancements to troll; however, there were extremely limited platforms on which to interact with victims. O’Sullivan and Heinonen (2008) outline the internet as a fast-evolving field, presenting unexplored areas and changing conditions for research analysis. However, there was a lack of research during the early stages of the *incubation period*, and this enabled trolling to emerge as an unforeseen symptom of the swift progression to unexplored interactive environments.

**Manifestation period (2006 to November 2011)**

Reciprocal functions of Web 2.0 redefined journalism and news transitioned from a linear process to a two-way communication model (O’Sullivan & Heinonen, 2008), in which readers could respond instantaneously, immersed in an interactive environment. This shift in the function of journalism enabled by new technology was called participatory journalism by Reich (in Singer et al, 2011) who spoke about user-generated content become part of the product that news organisations distributed. “Some journalists described comments as a necessary evil. Comments, they felt, were necessary to attract audiences and thus survive financially. But comments also were evil because their standards of expression were seen as intolerable” (p. 103). Although news organisations were adapting to the new world of participatory journalism that encouraged reader input, this process enabled negative feedback as well as positive. In this time-period, news environments were predominantly designed to engage readers more than traditional media sources had. Vine (2012) claimed newsrooms initially regarded interactivity as a beneficial technique to news distribution: “editors appear to be enthusiastic about social networking capabilities, and are embracing, or ready to embrace, interactivity such as that offered by Facebook and Twitter” (p. 170). However, they did not yet foresee the emergence of trolling and the impacts this would bring. News websites initially encouraged readers to use interactive features including ‘share’, ‘like’, ‘follow’ and ‘comment’ in response to their articles, enabling readers with similar interests to connect. Subsequently, this interactivity promoted and increased the readership of a given article. Falling advertising revenues have resulted in media organisations having to work harder to build loyal readerships and create an audience to respond to advertising, in
order to keep the business model afloat. New media tactics are, therefore, attractive to publishers due to their contribution to revenue. In 2014, Carey described the two-way communication model as having “high levels of perceived satisfaction” (p. 121) and links this to higher news readership retention rates:

Many news companies have recognised the potential of harnessing social media interactivity to increase exposure to news stories and create online following . . . many researchers have placed interactivity among the key components necessary to understand new media technologies. (Carey, 2014, p. 119)

At this time, interactivity was highlighted as a crucial factor promoting an article to its target audience. As a result, newsrooms monitored communications to structure and refine future articles to those best received by the public to ensure maximum readership exposure. The dynamic relationship that a comment stream enabled provided readers with a sense of importance and created a rapport between the reader and news provider. This continues, to some extent, to present day. It does this by allowing readers to express their opinions and to provoke conversation with other members of the public. This often prompts an exchange of ideas and/or public debate in an open online forum. In the manifestation period, this was the exact function from which trolling would emerge.

The manifestation period is identified as the time-period preceding the 2006 launch of Twitter and Facebook. Twitter was launched first and is a popular, user-friendly platform where online interaction could occur instantaneously on a global scale. Twitter engendered the required elements of technological advancements, coupled with the widespread public uptake, from which trolling stemmed and flourished. In this time-period, the term ‘trolling’ first emerged in Australian news media published by the SMH in July 2009, by News.com September 2009, followed by the ABC in February 2010. Trolling content within the manifestation period was sporadic, and the term was not yet comprehensively understood. Trolling began to appear concurrently in the Australian legal system in this time-period, with the first arrested Australian troll, Bradley Paul Hampson, charged in May 2010, and jailed in March 2010 for posting child porn on tribute pages for dead children. Disparate academic discussion on the issue emerged in
Social problem period (November 2011 to February 2014)

The social problem period emerged as a critical phase of the trolling evolution and established key themes for comprehensive analysis. The central factors that arise out of this time-period are interactivity, the gatekeeping role, viral news, and celebrity theory, which are further discussed and critiqued below.

Interactivity hinders the gatekeeping role of journalists

In this time-period, interactivity ruptured the journalist’s gatekeeping role, and this was the defining factor of the social problem period. The gatekeeping role is documented extensively in journalism education texts. Errington and Miragliotta (2011) outline the three main functions of journalism: to act as a “political watchdog”, “information provider”, and “facilitator of the public sphere” (p. 8). The third function is of particular interest in relation to the interactivity of new media where social media websites provided the platform for interaction between journalists and their audiences, thus enabling and encouraging unhindered conversations. In this time-period, interactive new media consolidated its role as a facilitator of the third function of journalism. Globalisation has connected people on a universal scale and academic discussions suggest interactivity is not a mere strength of new media but an essential part of its fundamental purpose (Errington & Miragliotta, 2011; Larsen, 2014). Comment streams are discussed in the newsrooms responses in chapter 4.6 and continue to play an active role in this function and are suggested as an important part of journalistic democracy.

Journalist Larsen (2014) discussed the growing extent to which this function was being implemented and how it was perceived by the public and workplace journalists: “the internet has democratised the art of critique . . . the removal of media and publishing gatekeepers has increased the number of diverse voices being heard and opened up more issues” (para. 3). Larsen (2014) claimed two-way communication models in relation to news distribution, such as comment streams, allow the public to publish unmonitored content. This is echoed in Suler’s
(2004) Online Disinhibition Effect research that suggested the “atmosphere and philosophy (of the ungoverned internet) contributes to the minimising of authority” (p. 324). With fewer, or no, trained journalists effectively performing their role as the gatekeepers, Larsen (2014) warned that this function is subject to abuse. Others claimed it has led to the rise of trolling itself, raising questions about the gatekeeping role. Domingo et al (2008) suggested issues such as trolling will inevitably arise if “the gatekeeper role is not maintained and enforced by professional routines and conventions that are said to guarantee quality and neutrality of institutional journalism” (p. 325), an assertion also supported by Reese and Ballinger (2001), Shoemaker (1991) and Singer (2005). However, as news moved online, the application of this function became difficult to enforce. The debate between industry professionals still exists about both the need for, and application of, the gatekeeper function. Steyn (2015) advocates for reducing journalist’s function as gatekeepers, citing Bill Gates saying, “the internet is becoming the town square for the global village of tomorrow” (para. 1), and highlights that the new global workplace should be embraced, not restricted. Cassidy (2006), however, notes that the “gate keeping theory...is certainly one of the most enduring in mass communication research” (p. 6) and suggests that it may continue to be so.

Gatekeeping is possible on websites that media companies own or control, where a news organisation owns their URL and to some extent their Facebook and Twitter accounts. The gatekeeping role is important, firstly, to reduce the risk of defamation, contempt of court, racial vilification, obscenity, and harassment as well as the other legal risks associated with being a publisher. Facebook has created a management tool a user may select, where prior consent is required for another person to post content on his or her own account. Most popular social media websites including Facebook and Twitter continually update their user functions and policies to counterbalance negative impacts of the interactive environment as they arise, a theme addressed in Questions D and E of this study. However, queries are raised here as to how much employers do control and should control what their employees say on social media. Conflicting opinions abound: “trolling with the intent to encourage active debate, without a pre-
determined malicious nature, is a healthy part of the public sphere’ (Funnell, 2012, p. 15); academic dialogue evaluates the importance of free speech counter arguments (Koerner, 2013; Savage, 2014) with Hylton (2014) saying, “freedom of speech does not mean freedom from consequences, just because you (may) have the legal right to speak your mind, does not entitle you to say whatever you want” (para. 1). However, in a media world with a nascent community of social media editors performing the role of gatekeepers attempting to manage characters wearing the masks of anonymity protesting about their rights to free speech, legal consequences were difficult to enforce. This topic is further unpacked in Question G of this study in chapter 4.7.

The emergence of viral anti-trolling campaigns
Here, amongst interactivity and functionality debates, the social problem period is identified as the time-period from November 2011 when trolling was thrust into the Australian news media headlines by feminist Sady Doyle’s (2011) #mencallmethings anti-trolling campaign (Griffin, 2011). Doyle originally published the campaign in 2009 on the feminist blog website Tiger Beatdown (Doyle, 2009) but received minimal media traction at that time (Gibson, 2011), falling in the manifestation period where only sporadic examples existed, and the term ‘trolling’ had not yet been identified within popular discourse as a social problem. However, in the social problem period two years later, the two key ingredients of timing and increased public interest were topical and the campaign went viral, pushing widespread criticism and debate into the community discourse of Australian living rooms, workplaces, and schools. Sady Doyle can be considered a pioneer of the anti-trolling revolution, who other notable public figures emulated, simultaneously challenging trolls by speaking out and initiating awareness campaigns.

Workplace and public backlash
In this social problem period, Australian radio host Kyle Sandilands was publicly accused of trolling and through intensive public backlash, lost a $2 million annual Holden sponsorship (Dixon & Waters, 2011). This is a clear example of trolling’s high social importance at this time. Amidst this period, a defining event occurred
in August 2012, when tv presenter Charlotte Dawson attempted suicide after a Twitter exchange with Monash University employee Tanya Heti, who told Dawson to "go hang yourself" (Lee, 2012). Heti was immediately suspended by her employer, demonstrating that employers and institutions were recognizing what was at the time called cyber-bullying, as a danger and as an unacceptable workplace behaviour. This key event demonstrated the sheer enormity of trolling impacts, which by now were universally recognised to spill from the virtual into the real world. This event resulted in global headline news, contributing to international awareness and discussion of the social problem. Shortly after, Dawson arose as the #SayNoToBullying campaign ambassador, publicly recognised as an Australian role model and trolling victim speaking out against cyber bullying and mental health (Connelly et al, 2012). The important role Dawson played in shaping the way trolling was discussed in the Australian news media is unpacked in the next few paragraphs, via contextualization in celebrity theory.

**Celebrity theory**

Globalisation of news media has transformed celebrities from admired personalities to powerful agents of social change, with extensive reach across socioeconomic, political, and cultural boundaries (Brown & Fraser, 2003). During the past 50 years, academic literature has observed that those who wield the greatest social influence have evolved from being traditional heroes such as political, religious, and military leaders, to being celebrities (in many western cultures) (Boorstin, 1961; Braudy, 1986; Campbell, 1988; Gamson, 1994). Widely documented links with advertising success infest universal celebrity theory, with mass media platforms enabling consumer awareness and encouraging favourable attitudes towards endorsed products (Atkin & Block, 1983; Freiden, 1984; Friedman & Friedman, 1979; Kamins, 1990; Tripp, Jensen, & Carlson, 1994). According to Monbiot (2016), when leveraged as “powerful means of persuasion” (Brown & Fraser, 2003, p. 99), actors and models now receive disproportionate attention, “capturing much of the space once occupied by people with their own ideas” (para. 8). Marketing literature refers to the “halo effect” around celebrities, where people have a very positive view of celebrities; and when products or
particular health practices are recommended or endorsed, the golden glow of the celebrity is transferred to the products or practice (Howard, 2017, para. 50). Economic literature recognises the concept of “signalling,” where celebrity endorsements are employed to clarify optimal products in a saturated marketplace, providing a cognitive shortcut to an otherwise very complicated decision (para. 51). Although a number of scholars addressed the influence of celebrities on culture from an historical-critical perspective (Boorstin, 1961; Braudy, 1986; Gamson, 1994), very few communication theorists and researchers have systematically explored the effects of celebrities on individual and social values, beliefs and behaviour (Brown & Fraser, 2003, p. 97), an emerging role of the celebrity in the digital age.

The cultural obsession with celebrity does not “lie quietly beside personal values, it takes their place” (Monbiot, 2016, para. 5). Emerging celebrities are becoming the dominant social leaders of our time, influencing prosocial and antisocial behavioural norms mirrored by and reflected in society (Xu, Reijnders & Kim, 2021). Now recognised by both profit and non-profit organisations (Brown & Fraser, 2003), the celebrity role has evolved from endorsed product promotion to social beliefs and behaviour influence. “When celebrities speak, it seems the world listens, even when it comes to personal and public health” (Howard, 2017, para. 1). A prominent, measurable example was Charlie Sheen's HIV-positive announcement on the 17th of November 2015. Research by Ayers et al (2016) found Sheen's public announcement corresponded with 1.25 million people searching HIV-related information on Google, the greatest number ever recorded in the United States, and sales of in-home HIV-testing kits simultaneously reaching record highs. Howard (2017) points out that although “the findings show a correlation, not a direct causation” (para. 17), this case study is notable in the annals of celebrity theory, as Sheen (2015) did not make an immediate call to action for the public to get tested or use condoms, and public health groups did not use him as a face to raise HIV awareness. However, this swiftly but organically manifested in social culture. By contrast, a similar HIV announcement by American basketball star Majic Johnson in 1991, when Internet use was limited and HIV was less familiar, only spurred questions and conversations, and no
measurable public action resulted. The notable discrepancies in the 1991 lack of action, and 2015 action in response to similar HIV announcements highlight the new celebrity role: to empower and facilitate social change, supported by the popularity and increased use of the Internet. In an additional example, Angelina Jolie’s 2013 double mastectomy announcement (Payne, 2013) appears to have had a positive influence on public health through breast cancer awareness, despite possibly influencing an uptick in expensive, unnecessary genetic testing for the rare BRCA1 gene (Howard, 2017, para. 20). Typically, public health advertising uses “top-down strategies” (para. 54) with the most effective messages believed to come from the experts at the top of the information hierarchy. However, both Sheen in 2015 and Jolie in 2013 are examples of organic media events, amongst others, that defy the “top-down strategies,” suggesting social messages are best received by the public when they come from the bottom, from people who are not health experts. Howard (2017) further suggests “when celebrities speak out about health-related issues, it resonates (with the public) because they are speaking as humans, with just as much vulnerability to health impacts as anyone else” (para. 53). While evidently not qualified as doctors, in recent years celebrities have managed to assert both positive and negative impacts, whether intentionally or not, on societal attitudes towards public health and discourse. Eight of the 528 articles in the thematic content analysis of this study mentioned Charlotte Dawson, who emerged as a key celebrity who pioneered the emergence of broad public awareness, contributed to discourse, and altered in a positive, exponential nature the shape of trolling awareness in the Australian news media.

Celebrities are often vital components of entertainment education campaigns. Since the objective of the entertainment education communication strategy is to increase knowledge of a personal or social need, influence attitudes, or change overt behaviour through entertainment (Singhal & Rogers, 1999) (Brown & Fraser, 2003, p. 108). Two days after Dawson’s attempted suicide in 2012, she became the ambassador of the #SayNoToBullying campaign, educating adolescents, and the wider community, about the dangers of online interactions and how to cope as a victim more effectively. Dawson was able to personalise a
private issue, often endured alone, and through speaking out was able to connect with the public beyond what the words of an educational campaign alone could (Connelly et al, 2012). Brown and Fraser (2003) suggest “a celebrity can become so closely identified with a social issue that the person and issue can become synonymous” (p. 100). Some celebrities become central to communication campaigns through their own personal tragedies or battles with disease, as in the case of basketball player Earvin “Magic” Johnson, actor Michael J. Fox, and cyclist Lance Armstrong (Brown & Fraser, 2003, p.109). Charlotte Dawson’s trolling is another example of this link. The successful collaboration between celebrities and educational campaigns has been a way to reduce the impact and occurrence of trolling. Anti-trolling campaigns, and the use of celebrities, will be further unpacked in Question C of this study in chapter 4.3.

Social epidemic period (February 2014 to June 2016)
Trolling was further thrust into the public spotlight after the February 2014 suicide of Charlotte Dawson as a result of trolling (Webb, 2014). Subsequently, trolling emerged as a ‘hot topic’ for discussion and buzzword in news headlines, opening conversations and debates about online cyber bullying. Although reports had begun to emerge including UNESCO researchers Barton and Storm’s (2014), Fernando (2014) and Friedersdorf (2014), the true scope of trolling-specific social impacts on Australian journalists were not yet comprehensively understood by the public. A social epidemic is defined as “a sudden, widespread occurrence of a particular phenomenon” (Levine Grater, 2009, para. 3). Dawson’s 2014 suicide was a catalyst for an exponential increase in trolling articles in the Australian news media, as demonstrated by the thematic content analysis raw data in this chapter. Trolling emerged as a buzzword with widespread use in popular discourse, but this also resulted in a diluted meaning. The understanding of the term evolved to a less specific association with any form of unwelcomed online interaction, such as persistent disagreement with political views, rather than referring to a deliberate and sustained online attack. Also in this period, the gendered nature of trolling became evident with reports by the Media Entertainment Arts Alliance (2016), Pew Research Centre (Duggan, 2017) and Gender Equity Victoria and Media Entertainment Arts Alliance (2019)
documenting women were trolled more than men. These conversations revealed the extremes of intense vitriol being shared online and subsequent victim impacts, and this further amplified the sense that trolling had become a social epidemic. Key events in this period include:

- In August 2014, the Gamergate saga sparked one of the most infamous and vicious trolling attacks ever reported, when in the context of the high pressure Silicon Valley game industry, Eron Gjon, the ex-boyfriend of gaming developer Zoe Quinn, posted her public information on his blog, which inadvertently started the gaming-industry culture war (Grow, 2015).
- In February 2015, Lindy West published an article in The Guardian, What happened when I confronted my cruellest troll, inspiring this research study (West, 2015).
- In June 2015, Clementine Ford posted a nude photo of herself with “Hey #Sunrise get fucked” written across her chest, accusing Sunrise of victim blaming. This prompted 20-year-old troll Ryan Hawkins to call her “lesbian scum” and threaten on Facebook to rape and bash her. His threats made international news and prompted discussions about victim impacts and potential legal ramifications as Ford pressed criminal charges (Ford, 2015).
- In August 2015, Former federal Labor leader Mark Latham resigned as a columnist for Fairfax Media’s Australian Financial Review after being outed for trolling several prominent Australian female journalists including Annabel Crabb, Mia Freedman, Leigh Sales and Lisa Wilkinson (Di Stefano, 2015).

**Saturation period (June 2016 to December 2017)**
The research identified the *saturation period* as commencing with Donald Trump’s 2016 election campaign and victory. During this year, the term *troll* flooded global news media, and the meaning of the term broadened. Trump used Twitter to an extent never seen before in an election of this scale, sharing what was often called out by others to be false information. In 2016, major publications headlines included The New Yorker’s: Trolls for Trump, The Guardian’s: Invasion of the troll
armies, from Russian Trump supporters to Turkish state stooges, and The Washington Post's: How the trolls won 2016. These were part of a tsunami of concerned journalism and commentary about the role internet contributors had played in democratic processes further echoed by the then Prime Minster Julia Gillard's intervention with social media companies that played a major role in opening up the debate publicly (Hunt, Evershed & Liu, 2016). However, increased media traction of trolling resulted in a dilution of the meaning of the term, and concern about electoral impacts diverted attention from victim impacts. The meaning of trolling shifted from a deliberate online attack on a person to a broader meaning including sharing fake news. However, victims continued to feel impacts prompting Jezebel writers such as Lindy West to leave Twitter declaring in January 2017, that Twitter is "unusable for anyone but trolls, robots and dictators" (West, 2017). In the same month, Australian journalist Jenny Noyes called for women to hold a Twitter strike in support of West (2017). This led to some journalists leaving the platform. The tsunami of coverage and dilution of meaning made it no longer feasible to continue the content analysis process used to this point without compromising the reliability of the analysis.
Chapter 4.2

Impacts

Results and Discussion

Responses to in-depth interviews with ten trolled journalists in 2015

*Question B - What are the impacts on trolled journalists in the Australian news media?*

The study provides some of the first comprehensive analysis of the dangers faced by many Australian journalists working in the news media. The study describes the types of violence and threats journalists often encounter daily and considers how these incidents affect their ability to conduct their journalism practice. Identified emerging trends from reported incidents reveal the extent of emotional and behavioural impacts, and suggestions aim to improve the ways in which the safety concerns of journalists are addressed.

The research design strategy applied mixed methods of thematic content analysis and in-depth interviews. A deliberate network sample was made of ten industry professionals, between the ages of twenty-three and sixty-five, five males and five females. Each had at least five years’ experience in the journalism field, self-selected as being a trolling victim and willingly offered their experiences. These participants were selected for semi-structured interviews of approximately an hour duration. The participants were asked questions about their personal experiences with trolling in the journalism industry. Participants reflected on any negative personal impacts, identified any workplace support systems in place and specified any personal training they received from university or their employers. Responses were then analysed for emerging themes in synthesis with respective theoretical frameworks for discussion and critique. As qualitative responses reached a point of saturation, ten victim impact interviewees were considered.
adequate to frame the personal and professional implications of trolling in the Australian news media.

4.2.1 In-depth interview results
Qualitative and quantitative data results from in-depth interviews indicate that all participants reported a level of distress after their trolling experience. However, the intensity of personal impacts varied between participants, indicated by differences in the wording and tone of individual responses. While responses to some questions were unanimous, responses to other questions were diverse, and emerging themes can be identified from the data to present a collective narrative of participants’ responses. The following findings display the study’s theoretical approach and group the interview responses within these emerging themes. Comments presented in this section have been edited for concision and to ensure anonymity.

The emerging themes identified from the in-depth interviews include: the prevalence of trolling, emotional impacts, behavioural impacts, newsroom impacts, and future impacts. These themes were reflected upon and linked within the study’s theoretical framework. The study found that feminist activists first spoke out about trolling to acknowledge the emotional and behavioural impacts and spotlight trolling as a significant social problem that required action from stakeholders at multiple levels. By 2016 when trolling was established as a social problem, the content analysis of 528 news articles from chapter 4.1 showed the focus of news content shifted from advocacy for change, to more systemic reporting of how stakeholder groups were responding. Although the feminist critical paradigm encompasses all themes because it identifies the overarching patriarchy, the infographic below marks the shift in the focus from feminist activism, to a systems approach as displayed below.
Participant responses to in-depth interview questions were grouped and presented within the thematic categories below. This study’s definition of trolling was outlined to participants before each interview commenced.

**Prevalence of trolling**

Asked if trolling had become prevalent in the Australian media industry, all ten participants indicated that it had. These are some of their responses.

- It's becoming a very big issue.
- I don't think it's probably right to accept that it's going to happen as people shouldn't do it, but it is going to happen.
- We cop quite a lot of abuse.
- Trolling itself has become a significant issue.
- It’s more prominent than ever before.
- It's part and parcel of the job.
- Trolling is evident everywhere you look really.

When asked how many times they had personally experienced trolling, all 10 participants noted the ubiquitous nature of trolling in the journalism industry.
Their experiences of trolling have been collated to provide quantitative data for analysis. Figure 11 below illustrates their range of answers, with comments highlighting the extent of trolling’s frequency, including:

- It’s almost too high to count.
- We cop a huge onslaught.

**Figure 11**

*Frequency of trolling comments received by participants*

<table>
<thead>
<tr>
<th>How many times have you been trolled?</th>
<th>n=10</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Once</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Tens of times</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>Hundreds of times</td>
<td>6</td>
<td>60</td>
</tr>
<tr>
<td>Thousands of times</td>
<td>1</td>
<td>10</td>
</tr>
</tbody>
</table>

**Trolling terminology**

As the trolling examples discussed in the literature review included political insults, rape threats and attacks based on appearance, participants were asked what trolls had said to or about them. Their responses about the nature of the comments or the target of the criticism are listed in Figure 12 and clustered into categories in Figure 13 below.

**Figure 12**

*Clustered trolling comments received by participants*

<table>
<thead>
<tr>
<th>Categorised trolling content</th>
<th>No of participants who mentioned each category</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appearance</td>
<td>5</td>
<td>50</td>
</tr>
<tr>
<td>Politics</td>
<td>4</td>
<td>40</td>
</tr>
<tr>
<td>Intelligence</td>
<td>4</td>
<td>40</td>
</tr>
<tr>
<td>Sexual promiscuity</td>
<td>3</td>
<td>30</td>
</tr>
<tr>
<td>Homophobia</td>
<td>3</td>
<td>30</td>
</tr>
<tr>
<td>Racism</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>Death threats</td>
<td>2</td>
<td>20</td>
</tr>
<tr>
<td>Rape threats</td>
<td>1</td>
<td>10</td>
</tr>
</tbody>
</table>
Figure 13

Specific trolling content received by participants

<table>
<thead>
<tr>
<th>Participant</th>
<th>Specific trolling terms and phrases received by participants coded into themes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>fucking black cunt, left tard, fucking left tard, hand wringing left tard,</td>
</tr>
<tr>
<td></td>
<td>uneducated cunt, poof, white racist, prick, child fucker, cunt, fuck you, toser,</td>
</tr>
<tr>
<td></td>
<td>get a life.</td>
</tr>
<tr>
<td>2</td>
<td>scrag, slut.</td>
</tr>
<tr>
<td>3</td>
<td>lefty scum, Tony Abbott hater, typical Fairfax bullshit, typical Fairfax journo,</td>
</tr>
<tr>
<td></td>
<td>Fairfax propaganda.</td>
</tr>
<tr>
<td>4</td>
<td>skinny, homophobic, poofter, gay, pseudo.</td>
</tr>
<tr>
<td>5</td>
<td>fat, old, ugly, you have the worst teeth I have ever seen, dirty feminist, full of</td>
</tr>
<tr>
<td></td>
<td>dumb shit, booze loving slut, dyke, lessos, homosexuals, quifer, gay, poofterism</td>
</tr>
<tr>
<td></td>
<td>should be illegal, dyke answer me this why do you always dress like men, do</td>
</tr>
<tr>
<td></td>
<td>you not understand this is how poofers and dykes coax children into their</td>
</tr>
<tr>
<td></td>
<td>perverted way of life, you should seek a cure for your illness, how can anyone</td>
</tr>
<tr>
<td></td>
<td>listen to a man who sticks his dick in another man’s ass, as my favourite singer</td>
</tr>
<tr>
<td></td>
<td>the late great Frank Sinatra said female journalists are no better than two dollar</td>
</tr>
<tr>
<td></td>
<td>hookers that about sums you up, I think faggots like you love to be offended, it’s</td>
</tr>
<tr>
<td></td>
<td>strange how the poofter Muslims are trying to take over however being a</td>
</tr>
<tr>
<td></td>
<td>poofter in Muslim countries receives the death penalty, I hope you die, all you</td>
</tr>
<tr>
<td></td>
<td>need is a good cock up you, you'll get my cock up you in a minute, you are</td>
</tr>
<tr>
<td></td>
<td>angling to get a root, delusional, mental illness, fake, cunt, misogynous, pathetic,</td>
</tr>
<tr>
<td></td>
<td>go get a real job.</td>
</tr>
<tr>
<td>6</td>
<td>slut, hoe, skank, worthless piece of shit.</td>
</tr>
<tr>
<td>7</td>
<td>fat, ugly.</td>
</tr>
<tr>
<td>8</td>
<td>do some work for a change you idiot.</td>
</tr>
<tr>
<td>9</td>
<td>lefty, idiot, retard, what would you know, get fucked, fuck head.</td>
</tr>
<tr>
<td>10</td>
<td>fat cunt, I hope you die, fuck you.</td>
</tr>
</tbody>
</table>

Trolling as a consequence of interactivity

All ten participants also linked the rise of trolling with the growth of online media. Comments that highlight this include:

- Trolling is like a bushfire that very easily runs out of control because of the social media component.
- Social media is used every day by almost everyone, so it’s quite prevalent.
- Social media has put us in direct contact with people we might describe as trolls.
- Trolls were out there before, sending nasty letters to newspapers that would never have seen the light of the day. We now, as journalists, have encountered them personally, almost daily, because of social media.
- It’s a massive change as there was no social media when I began.
- It is easier for people to do that these days as all they have to do is use a keyboard and they can remain anonymous.
Social media has enabled trolling to become more viable, and therefore more prominent.

Trolling has only existed with the implementation of social media.

Asked which media platforms were used to troll them, participants gave the answers illustrated in Figure 14 below.

**Figure 14**

*Platforms that participants received trolling content on*

<table>
<thead>
<tr>
<th>Answers</th>
<th>No of participants who mentioned each platform</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>Facebook</td>
<td>10</td>
<td>100</td>
</tr>
<tr>
<td>Comment Streams</td>
<td>9</td>
<td>90</td>
</tr>
<tr>
<td>Email</td>
<td>7</td>
<td>70</td>
</tr>
<tr>
<td>Telephone</td>
<td>3</td>
<td>30</td>
</tr>
</tbody>
</table>

**Emotional impacts of trolling**

As asked about specific instances of trolling and their emotional effects, all of the participants spoke freely, some cried as they related the stories, and all spoke with emotional pauses and inflection as they described the events and impact.

Comments illustrating a moderate distress level include:

- At this point I was fairly emotionally run out... someone suggested for me to inform the police, so I did.
- I do get really depressed about the state of humanity.
- In a personal sense, I felt a little affected by it.
- I know rationally that that hate is coming from an internal place of self-loathing, it’s not about me, it’s about them. They have a way at hitting you at your lowest, I guess, so that was shit.
- Here’s me, writing about Tony Abbott who is quite manly, and he’s doing a charity bike ride which is a very manly thing, so it was a rogue attack at my personal appearance, as I am myself quite petite in a way, I’m not a big guy, and you can see that in my profile picture, so people took it upon their selves to insult me for that.
• At the time I took it quite badly but in hindsight it probably wasn’t a huge deal.
• It’s difficult not to take it to heart, but you mustn’t.

Comments illustrating a high distress level include:
• This man has sent me rape threats before, this man has had a past history, and he’d been up in court for racist remarks, how horrific.
• I came home in tears that day.
• The emails were getting progressively more unpleasant, and I began to question whether or not I could handle it.
• I literally said fuck this and walked away from the business. I went away to the south coast and surfed every day. I still had to work every day, but I took some time to get my head around how to best deal with it and what we can do better.
• There was one who said I have the worst teeth I have ever seen. That really got to me, as I wasn’t really confident with my smile, I needed braces but never got any. I mean I’m a 40-year-old fairly confident woman who isn’t usually bothered by those things, but I went to the toilet and cried after that.
• It was so intense and such a huge onslaught; I couldn’t even follow most of it as I simply couldn’t keep up with the emails let alone the social media chain.

All ten respondents suggested they had become emotionally hardened over time indicating the development of a coping strategy as supported by research findings of Masullo et al (2020) and Martin and Murrell (2020). Some comments that illustrate this include:
• I’m used to getting attacked by The Australian or the conservative press almost every day now. I’ve become immune to it.
• I now have the opposite reaction [to feeling distressed], I actually enjoy it, I get a genuine kick when readers flip out about something I wrote.
• It neither shocks nor bothers me frankly, I think people make too much of a fuss about it.
• Not long term, we can all have a laugh about it.
• It hasn’t bothered me as much as others I know

**Behavioural impacts of trolling**
All participants had changed their behaviour in some way as a result of being trolled, suggesting current training and workplace support strategies are not adequate. Such changes are illustrated as follows:

**Personal behavioural changes**
- One participant joined the gym [after receiving insults about physique].
- One participant considered cosmetic surgery.

**Workplace behavioural changes**
- One participant started trolling the trolls themselves.
- Four participants stopped checking social media.
- One participant took two days off work after each of their articles is posted.
- One participant implemented a more caring and sensitive approach towards sources.
- One participant no longer enjoyed active debate online and no longer interacted with readers online.
- One participant had considered quitting.

**Resilience as a journalist**
All ten participants suggest the common advice of “don’t feed the trolls” is inadequate. Seven participants stressed the need to be resilient as a journalist. Comments that illustrate this include:
- Harden up.
- Don’t be too princess about it.
- You need be thick skinned.
- It’s easy enough to ignore it.
Two participants suggested that, if a journalist is going to respond, they should only engage with certain trolls. Comments that illustrate this include:

- You need to learn to pick your battles.
- It’s important to check their credibility before engaging in a critical debate. Check their Twitter account first, if they have no followers or have the little blue man as their profile, chances are they are a mindless troll, rather than an opinionated member of the public.

**Newsroom impacts of trolling**

**Trolling workplace training**

All participants mentioned that there were no workplace support systems in place for trolling specifically, with eight labelling this as inadequate. However, two participants, both under 25 years old, said trolling was “briefly mentioned” as part of their recent cadetship.

Comments illustrating that current workplace support strategies are not adequate include:

- I went and got drunk.
- I said fuck this and I left work for a month to go surfing.
- I don’t think anyone knows what the correct support system would be.
- Trolling needs to be addressed on case-by-case scenario.
- I’ve certainly had no training.
- Trolling training, haha, what is that?

**Trolling workplace support**

None of the ten respondents sought specific counselling, diagnosis, or medication in relation to their experiences, with only three participants being familiar with the Dart Centre for Journalism and Trauma, a resource centre that conducts seminars, training and support programs for journalists who cover violence, conflict, and tragedy around the world. The most common support systems utilised by all ten respondents were those of family, friends, and work colleagues.
Future impacts of trolling

Educational impacts

Of the seven participants who went to university, none received any training specifically in relation to trolling. Nine participants suggested that strategies for coping with trolling should be included within the university curriculum.

Two comments that raised the issue of training for trolling were:

• By saying that universities should implement more training would mean that I know what they currently do, which I don’t... maybe they need better viability for this stuff. Especially for those coming into a higher profile role, there should be some sort of preparation, however I don’t know how formal that needs to be.

• I’m not sure if it’s needed specifically, it’s the same with any bullying, and there is no one-size-fits-all answer.

Suggested future of trolling

All ten participants predicted that trolling is a part of the future of journalism. Six participants suggested trolling frequency and intensity would increase, with some comments, as follows:

• I see a bright and vibrant future (said with irony)
• It’s not going to die down anytime soon.
• It’s not going away.
• It feels as though it is getting worse.
• I think trolling will grow with the growth of social media.

Three participants suggested it is going to continue at the acknowledged current high frequency, with comments including:

• I don’t see it’s getting any worse than it is now, I feel this may be our peak; however, in the same breath, it’s not going to get any better any time some.
• I feel it’s probably going to stay the same, at least for the next while.
• Well, it’s not going to go away anytime soon, but at the same time it’s not new.
One participant suggested it would fluctuate based on the political climate at the time with the comment:

- I do think it will fluctuate. For example, we are about to go into a political campaign, so for political journalists it makes everything more heated, so it will really take off over that official 6 to 8-week period and then ease off. As a political journalist, it really depends on the political environment at the time.

One participant did not provide any definitive answer, and speculated with the comment:

- I can’t predict the future but again people with these crazy opinions have always existed.

4.2.2 Thematic analysis - a narrative of participants’ collective experiences

Prevalence of trolling

The collective experiences of participants were first observed from a functionalist perspective, and then were observed with the addition of the feminist critical theory perspective later. As a result, themes emerged from the data and are now discussed in this thematic analysis section. A functionalist perspective would argue that the anonymity provided by the internet presents a lack of negative consequences for a troll, which contributes to the vast prevalence of explicit trolling content online.

In 2015, when the interviews were conducted, examples of trolled Australian journalists had only recently begun to surface throughout mainstream media reporting as outlined in the literature review (on p. 24). Public acknowledgement of the rising problem spilled from newsroom dialogue to open public debate, asserting trolling as a prominent social issue. Although there was public consensus about its presence within the online realm, the impacts of trolling were a developing research area. Research by Stroud et al (2015) and Martin and Murrell (2020) investigated expectations of journalists to be “resilient”, as described by seven participants, and further discussed engagement with online perpetrators. However, with all ten participants labelling trolling as a “significant”
workplace problem and describing the suggestion to ignore trolls as an “inadequate” coping strategy, the scope of the problem begins to materialise from the data.

This is a crucial time-period in relation to the evolution of trolling because multiple factors coincided in a ‘perfect storm’. These factors include: the time-period, developing research, no universally recognised response, leading to trolled journalists feeling confused and overwhelmed. Participants were interviewed at an important time-period in the evolution of trolling in the Australian news media, with trolling having recently transitioned beyond the social problem period, into the social epidemic period as described in the emergence findings of this study. The thematic discussion section previously defined this social epidemic period as a time that reflects a surge in frequency of trolling related articles in the Australian news media. As the circulation of trolling related academic research was limited in 2015, and with seventy percent of this study’s participants experiencing trolling over one-hundred times individually, it is understandable that nine of the participants felt “overwhelmed”, which can be considered a fair reflection of the concerns of the greater community at this time. All ten participants described trolling as a “significant” and/or “problematic” workplace issue at the time, however, through subtleties in tone and phasing, two participants suggested that trolling had become an accepted by-product of Australian journalism practice, with Participant 3 reporting, “it’s part and parcel of the job.” The frustration with increasing exposure to trolling as part of their daily workplace environment, and a lack of refined knowledge of effective coping strategies meant many of the first journalists such as Lindy West (2015) were encouraged to speak out, writing and publishing their own personal experiences. This pushed trolling discourse into mainstream public debate.

**Anonymity provided by the internet**

While some online aggression is not anonymous and is explained by Rost et al (2016) who used social norm theory to explain why people “engage in online firestorms in order to defend higher order moral principles” (p. 24). They claim that non-anonymity helps to gain recognition, increases persuasive power, and
mobilises followers. However, they acknowledge the widely held perceptions that anonymity can exacerbate negative behaviour in online spaces. It is the position that anonymity does play a part that was affirmed by the participants in this study. One respondent emphasised the intensity of trolling provoked by its anonymous nature and used the metaphor for trolling of "a bushfire that very easily runs out of control." This argument holds that the lack of negative consequences for the troll intensifies trolling’s manifestation online.

Here, the two theoretical paradigms intersect and, in conjunction with one another, frame the narrative of participants’ collective experiences.

**Trolling terminology**
Observing trolling from a feminist perspective, patriarchal trends emerging from the data can begin to be unpacked. In a strategic network sample, half of the selected participants were female to allow gender comparison with theoretical reflection from the feminist paradigm. Distinct differences in the *ferocity* and *classification* of trolling examples were reported by participants of differing gender. All five female participants categorised their trolling experiences as sexually demeaning, with “slut”, “hoe”, “skank” and “fat” being some of the representative examples of trolling language, and one female participant further describing a threat of rape, echoing anecdotal examples of such threats in the literature review (Masullo Chen & Lu, 2019; Gardiner, 2018; Ryan, 2014; West, 2015). As distinct from accounts of patriarchal oppressive trolling by female participants, rape threats were not reported by any male participants, with three male participants reporting the political slur “left”, and two male participants reporting the sexual orientation slur “poof”. Although it is important to note that all trolling examples are deemed by this study as offensive, the data shows more descriptive and vulgar language of a sexually explicit nature was directed towards female participants.

**Emotional impacts**
The feminist and functionalist paradigms have both been considered when discussing the emotional impacts of victims. However, the increased prevalence
of trolling through interactivity and anonymity, as argued from the functionalist paradigm, suggests emotional impacts will still arise regardless of gender. The data reveals all five male, and five female participants reported a varied level of emotional impact, from moderate to severe. Of the three participants who reported high levels of distress, one was male, and disclosed, “I literally said fuck this, and walked away from the business, I’m going away, and I went to the south coast and surfed every day.” The data suggests emotional impacts of trolling are not gender-specific, and can affect victims in different ways, causing varying levels of distress.

While based on a small sample the findings show that all participants were distressed by being trolled, female participants expressed higher levels of distress linked with two crucial factors, an increased frequency of trolling coupled with severe and vulgar content. This confirms the applicability of the feminist paradigm. Findings suggest a combination of these two crucial factors provides the highest level of distress and were experienced by predominantly female participants. This high level of distress was evident in comments from two female participants who admitted to crying as a result of being trolled. The findings from these participants show emotional impacts were higher amongst females who experienced the two crucial factors: severity and high frequency.

The data further revealed all participants felt the need to become emotionally hardened over time as a coping strategy. Individual coping strategies suggested in the data led into the next emerging theme, the behavioural impacts.

**Behavioural impacts of trolling**

In this section, the study’s three theoretical paradigms begin to work in unison, with the addition of the third theoretical perspective, Ackoff’s systems theory paradigm. Ackoff’s system theory suggests social problems can be broken down into levels that can be individually addressed as part of the whole system. The behavioural changes of the interviewed trolled journalists, such as changing their physical appearance and removing themselves from online discussions, can be seen as a part of this system in the infographic in Figure 15 below which illustrates...
the differing levels of the trolling system identified in this study.

**Figure 15**

*Identified emerging themes within Ackoff’s systems model*

---

**Personal behavioural impacts**

All participants reported behavioural changes that in some way reduced interactions with readers as a result of being trolled. It is not only important to highlight the participants’ physical behavioural changes, such as “joining a gym”, but also their deeper underlying psychological changes. The psychological impacts emerged due to the personal nature of the in-depth-interview process. One participant was comfortable enough to divulge her battle with self-esteem as a direct impact of abuse categorised as appearance-based and revealed that she contemplated cosmetic surgery. In the sample group of 10, the data indicated that trolling categorised as appearance-based resulted in the highest level of personal change, equally affecting male and female participants, with two participants’ admissions supporting this analysis.
Workplace behavioural impacts

All participants reported a certain level of workplace behavioural change as a result of being trolled. Moderate workplace changes included one participant mentioning “they no longer enjoy active debate online” and “no longer interacted with readers online”. Significant workplace changes included four participants stating that they “no longer check their email”, a communication avenue many professions are required to utilise daily, journalists being no exception. One participant “took two days off work” and another participant “took a month off work” after being trolled, which are additional examples of trolling related distress adversely impacting upon journalism practice. The most serious workplace behavioural change in the study was noted when one participant further revealed they had “considered quitting” their job as a direct result of trolling. This study has identified that to cease practicing as a journalist due to the psychological and physical distress of trolling is considered the pinnacle or worst-case scenario and the most extreme potential workplace behavioural change as supported by Jane’s (2015) work on online harassment as a workplace health and safety issue. This systematic silencing of trolled journalists is labelled by Citron (2014) as hate crime and adds to the sense of alienation from society. This suppression of free speech, limited by decisions to “be more cautious”, “less vocal” and “hold back on arguments” in response to trolling impacts both career progression and the ability to play an equal part in society of affected journalists.

Seven participants reported the necessity of “resilience” as a workplace coping strategy and an expected industry norm for journalists as supported by Martin and Murrell (2020). When questioned on the effectiveness of resilience as a coping strategy, all participants questioned its longevity in withholding trolling-related distress. All participants labelled the commonly suggested advice “don’t feed the trolls” as inadequate.

Newsroom training and support

All participants indicated they had changed their behaviour in some way as a result of being trolled and would undertake further training and workplace support strategies if offered. Regarding support, all participants mentioned there
were “no workplace support systems in place for trolling” alone, with eight participants further labelling this as “inadequate”. With an overwhelming emphasis on the confusion brought about by having no universally agreed upon and recommended newsroom response to trolling, one participant revealed they “went and got drunk” as a coping strategy, with another reflecting his thoughts at the time of the trolling in the comment: “I said fuck this and I left work for a month to go surfing.” The most common support systems all ten respondents utilised were that of family, friends, and work colleagues. The in-depth interviews revealed that the support from family and friends was mostly on a personal and emotional level.

**Educational impacts**

At the time of interviews (2015), the weight of trolling behavioural and workplace impacts was not yet comprehensively understood in the industry, although researchers such as Phillips (2015) and Reagle (2015) were starting to reveal the extend of the issue. However, the significant need for the consideration and inclusion of trolling in journalism education was not reflected in the common texts used as a teaching basis for the multiple journalism courses taught in various Australian universities. Two of the youngest participants, both under 25 years old, said trolling was “briefly mentioned” as part of their recent cadetship, indicating trolling was beginning to be recognised as a workplace problem. However, a negligible amount of evidence-based research at the time provided effective coping strategies. All participants predicted that trolling would be part of the future of journalism, and nine participants suggested trolling should be included within current university courses as supported by Martin and Murrell (2021). To effectively include this in the curriculum, further research is required. This study responds to this call and provides data for educational institutions to draw upon when implementing a focus on trolling into their curriculum. Legal impacts, democracy and public debate are addressed in Question G in chapter 4.7 of this study (on p. 265).

Through analysis of interview transcripts, this chapter has described the impacts of trolling on ten Australian journalists, providing clarity about an important
stakeholder in the trolling paradigm. While the focus on individuals is important, the resolution to the trolling problem cannot be left to one level of the systems model, therefore other stakeholder groups need to be explored. The next chapter will investigate anti-trolling campaigns in the Australian news media.
Chapter 4.3 Campaign Responses

Results and Discussion

A chronological timeline of anti-trolling campaigns

Question C - How can the evolution of anti-trolling campaigns within the Australian news media be described?

Researchers have shown how unfettered free markets routinely manifest negative social phenomena, from the gender pay and promotion gap (Hanusch, 2013; Media Entertainment Arts Alliance, 2016), to the rise in casualisation in journalism employment (Jericho, 2018), which has particularly affected women (Ross, 2016), to patriarchy that tolerates abuse (Cusmano, 2018; Posetti et al, 2021). Gender-based abuse of women can be considered an inevitable aspect of the lack of regulation of digital communications platforms, as comments fuelled with hatred and vitriol drive traffic and attention to social media platforms and news websites (Phillips, 2015). This study describes how anonymity has exacerbated the trolling problem through the dehumanisation of internet users, but the underlying root cause is likely to be embedded far deeper within societal ideologies such as racism and other forms of discrimination including ageism and the bullying of children. These broader problems have many specific manifestations, including the trolling of journalists. Discussion of them provides context for this study’s analysis of the trolling of journalists. In context, it is understandable that the first anti-trolling campaigns that emerged to combat the abuse, both globally and in the Australian news media, were motivated by feminists. In an explorative process using grounded theory, this idea will be unpacked through thematic analysis, discussion, and critique in the following section.

The research design strategy involved mixed methods of thematic content
analysis of 528 trolling related news articles and archival evidence. Using Google as a search tool, the terms: “troll” + “campaign” were investigated with the location set Australia. From analysis of the first ten pages of the Google search results, totalling 100 articles, 13 significant campaigns were identified as archival evidence. Activist responses were plotted on a timeline that documents and illustrates the evolution of anti-trolling campaigns in the Australian news media. Commentary patterns emerging from thematic analysis of this data are compared with the evolution timeline and are presented in an additional infographic.

4.3.1 Qualitative data results
A chronological timeline of significant anti-trolling campaigns in the Australian news media was mapped to correspond with the periodical trolling phases identified in this study. Identified campaigns were divided into those based on sexism and those based on racism, and are displayed in Figure 16 on the following page.

It is important to highlight that the timeline used for comparison between the identified trolling time periods and the anti-trolling campaigns concludes at the end of 2017. The study recognised this time-period as the saturation period, due to the overwhelming amount of trolling-related content that obscured the other journalism-related trolling that is the subject of this study (as section 4.1 detailed). Prominent anti-trolling campaigns beyond 2017 were not excluded from the thematic discussion and will still be referenced as further supporting analysis, although they do not appear in the chronological timeline of Figure 16.
Figure 16

*Chronological timeline of evolution of anti-trolling campaigns in the Australian news media derived from chapter 4.3 data*
4.3.2 Qualitative and quantitative thematic analysis

Natural progression of social movement campaigns

The natural progression that many social problems follow shaped the path of the anti-trolling campaigns in the Australian news media. Put simply, there can be no opposition to a problem until the problem itself has been recognised. Social movements challenge social problems, therefore, the social problem of trolling must already be present in society, in both understanding and prevalence, for campaigns to be developed as a response strategy against it.

Reflection of society

Consensus among academics and industry professionals, also noted in Australian journalism education, suggests the “media as a reflection of society” (Bowman, 2014, p. 48). This study recognises this truism and further suggests that if the media is in fact a reflection of society, trolling may be considered a reflection of the negative parts of that society. Jane (2015) and Gorman (2019) echoed these sentiments and suggested the manifestation of technology-enabled abuse reveals some unpleasant ‘deep-rooted’ human behaviour. If this is true, then the racism, sexism and condoning of violence identified as themes from the study of the anti-trolling campaigns would appear to reflect the broader social issues in Australia over the time of this study.

Feminist campaigns

As suggested in the section 4.1 and 4.2 findings, the first trolling attacks in the Australian news media feminists were the first to identify the problem of trolling and to engender an anti-trolling or troll-oppositional discourse. Feminism as a movement actively provokes conversations concerning gender equality, and representatives are often the first to identify new emerging issues and raise awareness of their presence and impact in society. With feminism’s main intention to stimulate public awareness of gender imbalances that may not yet be recognised or universally understood by the general population, it is rational to propose this group would receive criticism from people who are content with or benefit from the social norms at the time. Feminists who rebuff the status quo and disrupt collective ideologies by questioning patriarchy or white male often receive
the first negative trolling comments, which this study's findings have shown can quickly escalate from critical debate to severe and extensive abuse in Chapter 4.1. The logical progression of widespread feminist-based trolling attacks led to the first anti-trolling campaigns being spearheaded by feminists.

The section 4.1 findings further suggest trolling in response to feminist writing or journalists who present at feminist are more frequent in the Australian news media. This trend is supported by the findings of this section and is reflected in the chronological timeline in Figure 16 (on p. 84), showing seven of the thirteen identified key anti-trolling campaigns are feminist-based. Multiple feminist trolling victims are referenced within each identified anti-trolling campaign, representing a collective group of females challenging the unacknowledged and social oversight of the widespread impacts, accentuating its prevalence at the time.

#Mencallmethings: the pivotal campaign

Sporadic trolling examples documented in the manifestation period began to plot the emergence of trolling and comprise the individual components that accumulate in trolling’s ‘perfect storm’. The ‘perfect storm’ was the point when mentions of trolling peaked, propelled by victims speaking out attracting increased media coverage and subsequent heightened public awareness. The philosophical concept of the episteme developed and articulated by Foucault (1971) in The Order of Things can be used to contextualise and explain the importance of the pivotal #mencallmethings campaign in shaping the evolution of trolling in the Australian news media. Foucault (1971) used the word episteme to describe the knowledge and thoughts possible in a given culture at a given time. He explains that epistemes can change over time (giving examples of how the 16th Century episteme gave way to the 17th Century episteme). This highlights the significance of timing and unconscious structures for a social problem to attain public recognition. Foucault noted that a “slow long effort to bring reflection upon” (p. 195) issues can cause evolution on the level of the episteme. Timing and changes in social construction of acceptable online behaviour are crucial within any given society for a problem to surface as a universally recognised issue. In the
identified *manifestation period* on November 7, 2006, Sady Doyle first published the #mencallmethings article calling for action on trolling; however, the article did not accomplish any traction within the media or public discourse. However, on November 20, 2011, the #mencallmethings article was republished with an explanatory foreword to provide context and went viral. From this it can be inferred that the social conditions and the social construction of trolling had changed to allow widespread publicity of an article/campaign that had failed five years prior. This significant milestone was identified by the study as the precise start of the *social problem period*. The campaign cemented trolling as a recognised social problem from which all other anti-trolling campaigns followed. The timeless characteristic of online publishing enables a given article to sit dormant online and resurface at a later date when epistemic conditions permit widespread mainstream media traction.

**#MeToo campaign: the shifting episteme**

While it is not specifically about trolling, the #MeToo campaign spread online and expressed dissatisfaction with patriarchal attitudes and behaviour. It is the broader context of the call out culture and new intolerance in which the anti-trolling campaigns emerged. The discussion of sexual harassment in the workplace is significant throughout feminist history and theory, and intersected with the campaigns identified in this study. Observing mainstream media from a feminist perspective requires reflection on the media representations of women. Redfern and Aune (2013) suggested the present “raunch culture” (p. 52) is fuelled by the media’s tolerance of (and profit from) the objectification of women, with *The Sun’s* page three topless women being a blatant example. After receiving widespread public criticism and 100,000 petition signatures, the misogynistic page three articles ceased on March 29, 2017. Further evidence of objectification is found throughout the descriptive vitriolic language aimed at against female journalists online. Redfern and Aune (2013) further state “women’s visibility in popular culture does not mean women are valued, safe from violence or equal” (p. 8), and exposed flaws in the preconceived ideology of a cohesive society. Gender inequality and workplace sexual harassment continue to emerge in the thematic content analysis of this study, and pattern-coded data can be compared against...
the #metoo movement for further reflection and discussion. Campaigns around a single issue have the potential to impact upon some people and can result in specific legislative changes. In addition, a study by Levy and Mattsson (2022) found the #MeToo movement had resulted in a higher propensity to report sex crimes. Also using data from the US, Szekeres, Shuman and Saguy (2020) found a decreased dismissal of sexual assault following #MeToo. This shows that campaigns in synergy create a more collective message and are more difficult for society to ignore. It is the cumulative effect that creates public discourse, provokes community debate, incites criticism and academic inquiry, and can result in an altered social tolerance or a shift in the episteme. Figure 17 on the following page displays the cumulative effect of the anti-trolling campaigns identified in this study that could have influenced social intolerance of workplace sexual harassment. Legal change can be seen as a milestone in the evolution of social intolerance for an issue. The #MeToo and associated movements prompted more widespread acceptance and use of the laws against online abuse as discussed in research by Atwater (2018) and Smith and Ortiz (2021).
Figure 17

The cumulative effect of the identified anti-trolling campaigns on shaping universally accepted social tolerance of workplace sexual harassment derived from chapter 4.3 data.
As suggested in Foucault’s (1971) notion of the episteme, this ‘perfect storm’ arose and enabled the eventual success and social uptake of the #MeToo movement. Although “Me Too” was first mentioned in 2006 by Tamara Burke to bring awareness of sexual abuse of women of colour on My Space, the ineffective two-way communication model, along with limited algorithm patterns of the previously popular social media platform, did not result in any notable media traction. The second identified key event in the #MeToo movement was in 2015 when Italian model Ambra Gutierrez first publicly reported Harvey Weinstein for sexual abuse; however, the media was quick to victim blame and portrayed Gutierrez as an opportunist. The third identified key event in the #MeToo movement was in 2016 when Ukraine journalist Anastasia Melnichenko commenced the #IAmNotAfraidToSpeak campaign, highlighting the unacceptable prominence of sexual harassment in the journalism workplace. The cumulative effect of the debate snowballed and escalated the #MeToo movement into one of the most influential feminist-led campaigns of the generation. On October 15, 2017, Alyssa Milano posted on Twitter following widespread accusations of predatory behaviour by Harvey Weinstein, encouraging others to share their own stories to shine the spotlight back on workplace harassment. Milano wrote, “if you have been sexually harassed or assaulted write #MeToo as a status, then we give people a sense of the magnitude of the problem.” On Twitter, #MeToo had been used more than 200,000 times in 24 hours and over 500,000 times in 48 hours (Sini, 2017). On Facebook, #MeToo was used by more than 4.7 million people in 12 million posts in the first day (Park, 2017), many of these were celebrities. The presence of celebrity (discussed in the emergence findings of section 4.1.3) further amplified media traction and public awareness. Coupled with Foucault’s (1971) episteme theory, the intricate nature and complexity of social structures must be considered when studying any substantial social change.
The influence of campaigns on cultural views

The persistence of feminist-led campaigns became the catalyst and repeatedly pressured workplace sexual harassment back into public discourse for debate. This continued persistence influenced the public to re-evaluate its standing on the issue. By applying thematic analysis, this study has identified three stages of social change in response to campaigns:

1. public awareness of the issue
2. cultural tolerance of the issue being questioned through exposure to victim examples and debate in mainstream discourse.

This evolves to

3. a new public intolerance of the issue and active distain for those who do not adhere to new ideology

A once commonly accepted social norm emerges as a newly outed, unacceptable social behaviour, embedded in the new belief system of that society. These developmental stages describe how societies mature and evolve, with examples of similar progression patterns including people of colour’s right to vote, people of colour’s right to own land, women’s right to vote, women’s right to drive in Dubai [2019], and recently, in Australia, those who identify as LGBTQI’s right to marry [2017], women’s right to legal abortion [2019], and those suffering from a terminal illness’s right to euthanasia [2019]. This cluster of issues fit with the notion of intersectionality, which has been often used by feminist scholars. It was first described by Crenshaw in 1989 to describe how overlapping systems of power, such as race, class, gender, and sexuality, marginalise groups (Cooper, 2016).

Australian racism culture in 2015

As some of the journalists interviewed in chapter 4.2 said they had been trolled with racist comments, it is relevant to discuss recent discourse about racism in Australia. As defined in the 2017 Cyber racism and community resilience strategies for combating online race hate by academics Andrew Jakubowicz, Kevin Dunn, Gail Mason, Yin Paradies, Ana-Maria Bliuc, Nasya Bahfen, Andre Oboler, Rosalie Atie and Karen Connelly, ‘race is quintessentially a social and cultural construct, drawing on local and global historically significant markers of difference to justify
unequal relations of power” (p. 21). This broad understanding of racism will be adopted by this study. Racism is a social issue, and as Ackoff’s (1974) systems theory suggested, “operates at many levels in the social ecology of modernity, from the broad panorama of global difference and power to the internal dynamics of cognition and identity in the individual” (p. 21). Race is “deeply embedded in psychological menus people draw on” (p. 21) when making conscious or subconscious decisions about their own beliefs, behaviours and opinions of others. By classifying and separating groups based on particular attributes such as ethnicity, religion, appearance or sexual preference, power imbalances have the potential to be leveraged against minority groups, and from this racism emerges. While Islam is a religion its expression in some cases connects with ethnic practices. Therefore, in this study, trolling based on Islamic identity is examined alongside racist trolling.

A clustered group of four anti-trolling campaigns that were categorised as anti-racist-based were identified in 2015 within the social epidemic period. From this clustered group, professions of trolling victims were identified to be different, and were documented to include: an Australian ambassador, a politician, an athlete and an academic. The anti-racist content of these anti-trolling campaigns was analysed to determine what racial groups the trolled victims identified with, to begin to analyse the motives behind the trolling abuse that the campaigns responded to. Race and religious based anti-trolling campaigns were further categorised into two themes: Indigenous-based and Islamic-based. Two anti-trolling campaigns were categorised into each category respectively and are displayed in Figure 18 on the following page.
Figure 18

A snapshot of the race and religious based anti-trolling campaigns identified in 2015

Note: The racism-based anti-trolling campaigns were identified in the thematic content analysis of this study. The snapshot displays the social epidemic period portion of the Chronological timeline of the evolution in the Australian news media derived from chapter 4.1 data in Figure 9 (on p. 51).

Indigenous-based trolling and Islamic-based trolling themes that emerged from Figure 18 (displayed above) prompted the further analysis and investigative discussion that follows.

**Indigenous-based racism**

Contemporary Indigenous-based racism in Australia emerged as the by-product of colonialism. Community attitudes have historically transformed slowly and unevenly, with the 1967 Referendum being the pivotal example in which 90.77%
of Australians overwhelmingly voted to amend the Constitution to allow the Commonwealth to make laws for Indigenous people and include them in the census (Parliament of Australia, 1967). “The significance of the 1967 Referendum has been somewhat obscured by a number of myths which include the misconceptions that the Referendum granted Indigenous people citizenship, the right to vote, wage equality and access to social security, among other things” (Thomas, 2017, para. 11). However, in terms of practical significance, its principal outcome was to raise the expectations of Indigenous rights and welfare in the Australian public sphere. Although subsequent legislation, including the Racial Discrimination Act (1975), the Commonwealth Racial Hatred Act (1995) and the Human Rights Equal Opportunity Commission Act (1986) (legal findings are discussed further in chapter 4.7 on p. 265) aimed to outlaw racial discrimination in the Australian community, the Australian political climate continues to tolerate racism. The apology to the Stolen Generation by then Prime Minister Kevin Rudd on February 13, 2008, remains a prominent example of transformational cultural views towards racial equality. However, the political ambiance at this time was not universal, with politician Peter Dutton boycotting the apology suggesting it would not deliver “tangible outcomes” (Flitton, 2018, para. 26) for Indigenous Australians. Subsequently he admitted he had overlooked the apology’s significance. Another recent example surfaced in 2019, when Pauline Hanson voiced her disrespectful attitude towards Indigenous culture, calling for Indigenous leaders to disregard their sacred beliefs and overturn a ban on climbing Uluru, likening the prevention of climbing the sacred rock to closing Bondi Beach (News Corp Australia, 2019a). News Corp Australia may have removed the name of the journalist who published this article as a preventative response to trolling. Consequently, the author cannot be referenced by name. Political journalist for The Guardian and Indigenous trolling victim Jack Latimore (2018) stated, “Australia is a far more racist country than its political class like to admit” (para. 2). He further claimed that the racism he observed “is no anomaly” (para. 8) and his observation supports this chapter’s findings. The animosity expressed by these prominent political figures lingers due to community leniency and has materialised as a plethora of racist abuse online.
Additional Indigenous-based anti-trolling campaigns identified in this study beyond the conclusion of the chronological timeline in December 2017 include the April 5, 2019, Melbourne Football Club Banner which read “We are taking a stand against trolls by tearing through tweets on our way to kick goals” (Lerner, 2019, para. 1). Broadcast on national television before the Australian Rules Football game commenced, the term troll was commonly used and understood, cementing itself in the Australian lexicon of 2019.

**Islamic-based racism**

Australia’s population reached 25.5 million people by the conclusion of 2019 (Australian Bureau of Statistics, 2019), and as the population continues to grow, it is also becoming more culturally diverse. The most recent 2016 Census report can be analysed to establish demographic trends at the time. The Census data verified Australia’s cultural diversity in 2016, with a population derived from more than 190 different countries, who spoke more than 300 different languages (Jeyaratnam, 2018). In response to the Census question “Is it a good thing for a society to be made up of different cultures?”, 80.4% of the population agreed, 4.4% of disagreed, and 15.3% neither agreed nor disagreed. However, difficulties arise with the practical implementation of diversity within Australian society throughout subsequent questions of the report, that highlight the complex nature of the issue. In response to the Census question “should all migrants be accepted regardless of where they came from?”, 53.3% of the population agreed, 23.4% disagreed, and 23.3% neither agreed nor disagreed. This data suggests a conditional acceptance of some cultures but not others, and possibly indicates the subsection of society from where the racism discussion emerges. Attitudes about which cultures have the right to be accepted within Australian culture, and furthermore, who has the right to determine which cultural groups are accepted or not, would arise within this conservative group. In response to the Census question, “should people from racial, ethnic, cultural and religious minority groups behave more like mainstream Australians?”, 48.7% of the population agreed, 18.5% disagreed, and 32.8% neither agreed or disagreed. This data further revealed the depth of the cultural diversity conflict, but most troubling was that the report found 32% of survey respondents had “negative” feelings towards
Muslim Australians, and 22% said they had “negative” feelings towards Australians of Middle-Eastern heritage (Jeyaratnam, 2018). Although anti-Islamic sentiments are seen as unacceptable by some Australian as evidenced by #IstandwithMariam (Stewart, 2015), examples can still be found in mainstream media and are voiced by some prominent Australian public figures.

The 2019 Australian political climate continues to tolerate Islamophobia when voicing right-wing attitudes of politicians Tony Abbott, Peter Dutton and Pauline Hanson. In 2015, Islamic leaders were furious at Tony Abbott’s suggestion that “the community does not do enough to stamp out extremism” (Medhora & Safi, 2015), accusing the then Prime Minister of promoting hatred and inflaming racism. In 2016, Peter Dutton said, “it was a mistake bringing Lebanese refugees to Australia” (Burton-Bradley, 2017, para. 1) and suggested Australia’s immigration policy should exclude those from Lebanese-Muslim backgrounds, linking the ethnic group with terrorism-related offences. The comments angered many. Pauline Hanson’s continued racist comments began in her 1996 maiden speech to Parliament, when she called for the abolition of multiculturalism and claimed, “we are in danger of being swamped by Asians” (Hanson, 2016, para. 6). More recently, in 2016, Hanson claimed, “now we are in danger of being swamped by Muslims” (para. 6), pushing for an immigration ban, and she controversially said, “go back to where you came from” (para. 1). In 2017, Hanson attended the Federal Senate dressed in a black burqa in an attempted stunt to support calls to ban the Islamic outfit. The then-Attorney General George Brandis reprimanded Hanson and rebuked her sharply for causing offence to a faith community. Brandis stated, “to ridicule that community, to drive it into a corner, to mock its religious garments, is an appalling thing to do, and I would ask you to reflect on your behaviour” (Murphy, 2017) and received a standing ovation from Parliament for doing so. This public reprimand suggested an intensifying refusal of racism and Islamaphobia. In support of the growing rejection of and dissent against the continued racist behaviour of some Australian politicians, Latimore (2018) suggested the need for accountability for their actions and further stated, “there can be no reconciliation until politicians acknowledge the harmful impacts of race baiting” (para. 1). Political journalist Katherine Murphy (2016) further suggested...
the media’s role in stamping out racism is crucial, and proposed that the continued presence of Hanson’s views within mainstream media normalised extreme ideologies and had the potential to render racism acceptable, “comprehending Pauline is not the challenge, engaging constructively with Hansonism is” (para. 1). The effect on the public of constant exposure to extremist views such as Hanson’s needs to be monitored by the media. Chopra (2018) suggested, “the prospect of PM Peter Dutton should be our wake-up call” (para. 1) and suggested the need for journalists to critique politicians in the media in order to keep society’s moral compass in check on key issues. The extensive anti-islamic trolling that targeted journalist and former Muslim Australian of the Year Susan Carland from 2015 to 2016 was a prime example, with Carland saying “social media, in particular Twitter, was toxic for women and Muslims” (Mannix, 2015, para. 4).

While vilification has been legally acknowledged since the 1980s with the integration of 18C into Australian legislation, racist trolling has seeped into global public debate, and although once considered an unfortunate consequence of the freedom Web 2.0 provided, public concern has shifted, and media coverage has magnified the understanding of racism-based impacts. Although the prevalence of racism within Australian culture is widely reported throughout news media, legal pathways towards prosecution or compensation are not universally understood, and incidences are often left unprosecuted. Legal responses to trolling will be discussed in further detail in chapter 4.7 (on p. 265). While the high court decision in the defamation case Fairfax Media Publications Pty Ltd v Voller (2021) 392 ALR 540 may reduce trolling that vilifies based on race, these effects are yet to be documented. This High Court case arose from vilification of Dylan Voller (Fairfax Media Publications Pty Ltd, Nationwide News Pty Ltd, Australian News Channel Pty Ltd v Voller (2020) NSWCA 102; (2021) HCA 27), the subject of a news articles published on Facebook, by several prominent news media organisations.

Protecting children from trolling: part of shifting the episteme
While cyberbullying has not been discussed in detail in this thesis, its trajectory is similar to that of trolling (Jeffrey, Peltier & Vannest, 2020) and some anti-cyberbullying campaigns worked in unison with anti-trolling campaigns to shift
attitudes about the acceptability and harmlessness of online abuse. Legal and policy concerns about trolling/cyberbullying arose first with children, who are viewed as one of our most vulnerable social groups. As a result cyberbullying education campaigns developed in the early 2010s (Foody, Muthanna & Carlbring, 2015). While the chronological timeline in Figure 16 (on p. 84) concluded in 2017, in 2018 two campaigns called ‘Dolly’s Law’ and ‘Carly’s Law’ received considerable media attention. Through thematic analysis protecting children from trolling emerged as a key theme for discussion. Amy “Dolly” Everett became the ambassador for the Australian outback hat company Akubra at 8 years of age, but on January 3, 2018, at 14, committed suicide after suffering continued online bullying. The teenage suicide became global news and Dolly became the recognised face of trolling’s greatest potential impact: the loss of a child’s life. The event became the final catalyst that forced the Australian Government to act, leading to research funding and crucial law reform known as “Dolly’s Law,” discussed in chapter 4.7 on legal responses (on p. 265). While the High Court decision of September 2021 may reduce trolling that vilifies, these affects are yet to be documented. In further supporting efforts to reduce child-based trolling, subsequent campaigns arose that called to ban smart-phone use in schools (Zimmerman & Bita, 2018) and called for a criminalisation of trolling that would include executing a social media order on perpetrators that would ban them from using social media websites (The Daily Telegraph, 2018).

While cyberbullying campaigns achieved rapid legal reform for children, anti-trolling campaigns did not. However, anti-trolling campaigns were able to support victims by breaking down a sense of isolation and encouraging them to speak about their experiences. This accords with the research Levy and Mattsson (2022) and Szekeres, Shuman and Saguy (2020) who documented effects of the #MeToo movement. Measurable action from all social levels, working in unison, is a practical application of Ackoff’s (1974) systems theory, but sadly it was only employed after a child’s life was lost. A similar example of a delay in a synthesised approach is evident in the progression of Carly’s Law. In 2007, 15-year-old Carly Ryan was murdered after being groomed online by predator Garry Francis Newman, who was later sentenced to life in prison. The story became global
headline news and Carly's mother, Sonya Ryan, initially began campaigning in her community and local news, which then progressed online and by 2017 received international news coverage, to raise awareness of online safety and for stronger laws to protect children online. Ten years later, on June 15, 2017, Carly’s Law was passed in Federal Parliament and a new offense made it a crime to plan to harm a child under 16 and, in particular, targeted predators who misrepresent their age. This provided police with the power to intervene before a predator has the chance to act and before a child is harmed. Foucault’s (1971) theory of the episteme can be applied to explain the time lag of law reform with Carly's Law, implemented after ten years of campaigning in 2017, as opposed to Dolly's Law implemented after 11 months in 2018. The ideal conditions for social change arose due to an accumulation of factors including:

- a cumulative intolerance to trolling after the 2015 suicide of Charlotte Dawson
- the study's identified saturation period (described on p. 62) when discussion pertaining to frequent and severe trolling content was abundant in public discourse and,
- the 2018 suicide of 14-year-old Amy “Dolly” Everett.

The identification of trolling as an escalating social problem pushed the issue to breaking point, and all levels of Ackoff’s (1974) systems theory responded. Applied in unison with Foucault’s (1971) theory of the episteme, the ideal timing for social change resulted.

In both examples mentioned, the Government did not respond until a child’s life was lost. Inaction is not an adequate contribution to prevention strategies. For effective solutions to emerging social problems, faster recognition and mitigation strategies are necessary, before life is lost, young or old. The loss of a child’s life should not need to be the catalyst for change. Academics, researchers, the government, and the public need to identify potential problems and their corresponding solution strategies quicker in order to generate a multi-level response before an issue escalates to a level where any life is lost as a result.

This chapter has outlined 13 significant anti-trolling campaigns providing clarity...
about action taken at the support level of the trolling paradigm. These actions include calls for legal changes that in some cases were successful, shifting epistemic understanding about the harms associated with online abuse, and identifying the issue. The first steps towards solving a wicked problem are naming and acknowledging the problem (Camillus, 2008), which the anti-trolling campaigns facilitate. Multiple stakeholders can actively influence trolling at different levels of the systems model, and separate analysis is important to provide a grounded view in line with Ackoff’s (1974) assertion that social problems involve impacts and reactions on multiple levels. The next chapter will investigate Twitter’s responses to trolling. Due to the scope of this study, it was not possible to analyse every social media platform used to distribute published articles by journalists and news organisations in Australia. The study, therefore, selected two social media platforms as the focus of further investigation, analysis, discussion, and critique. The chapter 4.2 findings from interviews with 10 trolled journalists (as described on p. 64) revealed that Twitter and Facebook were the two predominant media on which trolling content was received. These findings were further supported by documented examples from trolling victims within the chapter 4.1 findings (as described on p. 39) and through systemic calls for action directed at the two companies within the chapter 4.3 findings (as described on p. 82).
Chapter 4.4
Twitter Responses

Results and Discussion

Thematic critical analysis of Twitter responses

Question D - What policy and feature changes were implemented by Twitter in response to trolling from its launch in 2006 to 2019?

This study has documented widespread frustration with Twitter’s inconsistent enforcement of policies that ban abusive content on its service, and subsequent refusal to publicly explain or account for its actions and decisions. Despite multiple requests from academics (see, for example, Phillips, 2015; Jane, 2015; Citron, 2014; Barnes, 2018), Twitter has refused to release any data about the frequency of trolling on the service, how (or if) the company responds to reported abuse and, furthermore, what training procedures are in place for moderators. The adoption of this measured approach obscures the extent of the problem and critics would like Twitter to be more open to independent analysis of its moderation policy and processes. Although Twitter’s 2021 transparency report (Twitter, 2021) gives some insight about its removal of violating content including hate speech, abuse and harassment, its refusal to provide full operational transparency has fuelled growing discontent amongst users and the public and put pressure on the social media platform to respond before users cease use of the service (West, 2015; Quinn, 2021; Stevenson & Howcroft, 2021). By 2021, Twitter had been more vocal about introducing anti-trolling features and strategies than any of the social media platforms (Milmo, 2021), however, it could be far more open about its moderation processes. Collecting and independently analysing data is the first step in establishing Twitter’s accountability for failing to protect the human rights of its users online. This is relevant to the trolling of journalists in Australia because many Australian journalists were, during the period of this study, expected to use the platform for both work and personal branding. This
study identifies archival evidence of Twitter’s responses to trolling and collates this limited amount of accessible data as part of the first step towards academic analysis and critique of Twitter’s response.

Highlighting the continuing role social media will play in the future of journalism, a 2020 Muck Rack survey found that 85% of journalists considered Twitter the most valuable social network (Muck Rack, 2020). In addition, the Pew Research Centre found in 2017 that 58 per cent of internet users who have experienced online harassment said their most recent experience occurred on a social networking site or app, such as Twitter or Facebook, both platforms which journalists are expected to engage with as part of their professional practice (Duggan, 2017), while the second most common setting, at 23 per cent, was in the comment section of websites. Therefore, the study selected Twitter and Facebook as the two social media platforms that are independently addressed in Chapter 4.4 and Chapter 4.5 of this study. The Digital News Report 2021 (Parks et al, 2021) confirmed that Facebook and Twitter were the two social media platforms most used for news access with YouTube coming in third place. Because of scope limitations, YouTube was not analysed in this study.

Twitter and Facebook were addressed with separate research questions as their individual response strategies to trolling differed in several areas, including management acknowledgement, response times, number of strategies, type of strategies, depth of strategies and initiating research. The two research questions are indicated below:

1. Twitter responses to trolling in the Australian news media were addressed with Question D: What policy and feature changes were implemented by Twitter in response to trolling from its launch in 2006 to 2019? Results and discussions are displayed in Chapter 4.4 on the pages that follow.

2. Facebook responses to trolling in the Australian news media was addressed with Question E: What supportive strategies, operational modifications and corporate reform were implemented by Facebook in
response to trolling? Results and discussions are displayed in the 4.5 Facebook findings of this study (on p. 139).

The research design strategy which addresses Question D: What policy and feature changes were implemented by Twitter in response to trolling from its launch in 2006 to 2019? applies mixed methods of thematic content analysis and archival evidence. The study used collected data from chapter 4.1 (on p. 39) as a starting point. Of the 528 articles in the thematic content analysis, 17 mentioned Twitter's response to trolling on its platform in some capacity and were deemed relevant for inclusion. Of these, five further discussed or made commented on aspects of particular response strategies. To obtain further data, Twitter's website was explored, and relevant information was retrieved from three sections: About Us, Blog and Help Centre where its policies are posted. On each of these pages the search function was used to investigate the terms: “troll” “bully” and “abuse”. This led to a further 16 relevant articles which were included for analysis. Although Google was also used as a search tool with the search inquiry: “troll” + “Facebook” + “response” and retrieved 3,120,000 results, on closer analysis of the first ten pages totalling 100 articles, only one additional response strategy was found. Thematic analysis of the 36 articles retrieved as archival evidence grouped Twitter responses into three categories: policy responses, feature responses and managerial responses, which are independently documented chronologically and explored in the sections that follow. Emerging commentary patterns were integrated with the infographic of the evolution of trolling in the Australian news media for comparison, in-depth discussion and critique.

**Defining terminology**
This study adopts the following definition of the term policy: “a deliberate system of principles to guide decisions and achieve rational outcomes” (Klewes, Popp & Rost-Hein, 2017, p. 57). Policies are adopted by social media platforms, such as Facebook and Twitter, to outline the organisation’s statement of intent that governs the protocols and procedures implemented in both the Terms of Service and Rules of Engagement on the platform. However, policy does differ from rules or law. “Terms of service represent a non-legal avenue of recourse for professional
communicators wanting redress against trolls and troublesome users who might be damaging their brands and reputations on social media” (Pearson & Polden, 2015, p. 87). Although exact terms of service vary between platforms, policies give each social media company the right to suspend user accounts for what it deems to be misuse or misbehaviour (p. 87). This study defines the term feature as a distinctive attribute or aspect of a good or service that sets it apart from similar items (“Feature,” n.d.). Social media platforms use technological features that determine how the applications function. Over time, Facebook and Twitter have adapted their user interface features to reinforce company policies and enhance both the usability and safety of users on their platforms (Hutchinson, 2017).

4.4.1 Qualitative and quantitative data results
Clustered groups of responses emerged from thematic analysis applied to management responses (when management made public statements), policy responses (when adaptions of policy were announced), and feature responses (when features of the platform were changed) and were categorised into twelve chronological stages, and are listed below.

**Stage one:** Twitter’s first acknowledgement that trolling occurs on its platforms.

**Stage two:** Twitter declares trolling as a continuing problem.

**Stage three:** The reporting method for abusive content is streamlined.

**Stage four:** CEO declares responsibility for trolling on the service.

**Stage five:** A phone number verification is required for all users.

**Stage six:** The quality filter feature is implemented.

**Stage seven:** New accounts filtered for previously flagged content.

**Stage eight:** The Trust & Safety Council is introduced.

**Stage nine:** Accounts that promote violence are prohibited.

**Stage ten:** CEO declares public accountability for progress.

**Stage eleven:** Flagged accounts are muted.

**Stage twelve:** Technology experts Smyte are introduced.
The twelve chronological stages with their respective management, policy and/or feature responses are outlined in detail below. These descriptions are drawn upon as qualitative data for further thematic analysis, discussion, and critique in this section.

**Stage one: Twitter’s first acknowledgement that trolling occurs on its platforms.**

- **Policy response** *(September 14, 2012)*

From this study’s functionalist analysis, it is important to consider why consequences arise, and three days after *The Daily Telegraph* launched its #StopTheTrolls campaign Twitter made its first policy trolling response. Twitter amendment to their Rules of engagement, which stated, “technical abuse and user abuse is not tolerated on Twitter.com, and will result in permanent suspension” *(Twitter, 2012, para. 1).* This 2012 statement marked significant progress in six years, after the first Twitter Rules in 2006 had stated, “we do not actively monitor and will not censor user content” *(Twitter, 2006, para. 1).* The first Twitter Rules were short, consisting of 568 words in total, with one sentence to describe appropriate behaviours that covered the following policies: Impersonation, Privacy, Violence and Threats, Copyright, Unlawful Use, Serial Accounts, Name Squatting, Malware/Phishing, Spam, and Pornography. Although extreme examples of trolling may have been covered within these initial rules, it was not until the abuse and harassment policy was added that trolling was able to be addressed.

- **Management response** *(September 14, 2012)*

The policy update was accompanied by Twitter’s first management response that directly addressed trolling, again in response to accumulating pressure surrounding the #StopTheTrolls campaign that launched three days prior coupled with the Federal Government’s calls for Twitter to establish an Australian outpost to deal with community concerns *(Hildebrand & Matheson, 2012).* In a statement to the *The Daily Telegraph*, Twitter spokeswoman Rachel Bremer said, “we take
this issue extremely seriously and have recently updated our abuse policy to make targeted abuse against our terms of service” (Hildebrand & Matheson, 2012, para. 3). The findings revealed this policy ‘update’ was, in fact, a new policy that was not mentioned in the prior Twitter Rules.

**Stage two: Twitter declared trolling as a continuing problem.**

- **Management response**
  **(August 15, 2014)**
  Twitter’s vice-president of trust and safety Del Harvey responded to the trolling of Zelda Williams, the daughter of actor Robin Williams, with the statement, “we will not tolerate abuse” (Twitter, 2014, para. 3).

- **Policy response**
  **(August 15, 2014)**
  Twitter stated it would expand policies regarding self-harm and private information and improve support for family members of deceased users (Tsukayama, 2014).

**Stage three: The reporting method for abusive content is streamlined.**

- **Feature response**
  **(December 2, 2014)**
  Twitter’s updated features streamlined its abuse reporting method, allowed users who were not directly targeted by the abuse to have the ability to report it, added a ‘blocked accounts’ page that allowed users to view and edit accounts they had blocked, and prevented blocked users from viewing the profiles of the people who blocked them.

**Stage four: CEO declared responsibility for trolling on the service.**

- **Management response**
  **(February 2, 2015)**
  In direct response to Lindy West’s (2015) article, *What happened when I confronted my troll cruelest troll*, published in *The Guardian* earlier on this day,
Twitter’s CEO Dick Costello admitted the platform had failed trolling victims on his Twitter account:

We suck at dealing with abuse and trolls on the platform and we’ve sucked at it for years. It’s no secret and the rest of the world talks about it every day. We lose core user after core user by not addressing simple trolling issues that they face every day. I’m frankly ashamed of how poorly we’ve dealt with this issue during my tenure as CEO. It’s absurd. There’s no excuse for it. I take full responsibility for not being more aggressive on this front. It’s nobody else’s fault but mine, and it’s embarrassing. We’re going to start kicking these people off right and left and making sure that when they issue their ridiculous attacks, nobody hears them. Everybody on the leadership team knows this is vital.

(Costello, 2015, @dickc Twitter)

- **Management response**

(February 3, 2015)

One day after his statement, Costello stated he was “ashamed” (McGrath, 2015, para. 1) of how the company handled trolling and declared personal responsibility for its manifestation on the platform, further stating on his Twitter account:

Let me be very very clear about my response here. I take PERSONAL responsibility for our failure to deal with this (trolling) as a company. I thought I did that in my note, so let me reiterate what I said, which is that I take personal responsibility for this. I specifically said, ‘it’s nobody’s fault but mine’.

(Costello, 2015a, @dickc Twitter)

and added in a subsequent post:

We have to be able to tell each other the truth, and the truth that everybody in the world knows is that we have not effectively dealt with this problem even remotely to the degree we should have by now, and that’s on me and nobody else. So now we’re going to fix it, and I’m going to take full responsibility for making sure that the people working night and day on this have the resources they need to address the issue, that there are clear lines of responsibility and accountability, and that we don’t equivocate in our decisions and choices.

(Costello, 2015a, @dickc Twitter)
Stage five: A phone number verification is required for all users.

- Feature response

(March 13, 2015)

Twitter introduced a new verification method to combat troll accounts. A phone number verification was now required for all new accounts. Previously, all that was required was an email address, which could be obtained with no verification at all.

Stage six: The quality filter feature is implemented

- Feature response

(March 23, 2015)

Twitter launched a ‘quality filter’ aimed to “remove all tweets from users’ notification timelines that contain threats, offensive or abuse language, duplicate content, or are sent from suspicious accounts” (Welch, 2015, para. 1) and is displayed in Figure 19 below.

Figure 19

Twitter’s quality filter feature for users

**Stage seven:** New accounts filtered for previously flagged content.

- **Policy response**
  
  *(April 22, 2015)*

  Twitter revised a former policy that prohibited “direct specific threats of violence against others” adding the phrase “or promot[ing] violence against others”. The amendment meant any user who shared another user’s threat was now infringing the Twitter policy. This updated policy aimed to eliminate the loophole used by trolls who circulated other users’ threatening comments (Doshi, 2015).

- **Feature response**
  
  *(April 22, 2015)*

  In addition to the aforementioned policy change on the same day, Twitter simultaneously announced a new feature to help identify abusive content and limit its potential reach. The feature analysed the age of a Twitter account and compared the similarity of content between previously reported threats. This feature was designed to stop trolls immediately setting up another account when they have been previously reported and blocked. This new feature allowed Twitter to freeze an account until the user deleted the flagged tweet.

**Stage eight:** The Trust & Safety Council is introduced.

- **Management response**
  
  *(February 10, 2016)*

  Twitter introduced the Trust & Safety Council to deal with harassment and cyber bullying on the platform. The council was initially made up of 40 organisations around the world, four of which were Australian-based (Ockenden, 2016).

**Stage nine:** Accounts that promote violence are prohibited.

- **Policy response**
  
  *(November 18, 2017)*

  Twitter updated its policies to reduce hateful conduct and abusive behaviour. The first policy change was a new rule against violence and physical harm. This ensured specific threats of violence or calls for serious physical harm, death, or
disease inflicted on an individual or group would now be in violation of Twitter’s policies.

The second change expanded on Twitter’s Hateful Conduct policy and rules against abusive behaviour to now include rules to prohibit users promoting violence against other users, directly attacking or threatening other users on the basis of their group characteristics, as well as engaging in abusive behaviour that harasses, intimidates, or uses fear to silence another person’s voice. Twitter updated its rule against abusive behaviour to state:

Any account that abuses or threatens others through their profile information, including their username, display name, or profile bio. If an account’s profile information includes a violent threat or multiple slurs, epithets, racist or sexist tropes, incites fear, or reduces someone to less than human, it will be permanently suspended.

(Twitter Safety, 2017, para. 3)

In a further update promoted on its website’s blog as an active strategy against trolling, Twitter expanded its hateful conduct policy to state:

Hateful imagery will now be considered sensitive media under our media policy. We consider hateful imagery to be logos, symbols, or images whose purpose is to promote hostility and malice against others based on their race, religion, disability, sexual orientation, or ethnicity/national origin. If this type of content appears in header or profile images, we will now accept profile-level reports and require account owners to remove any violating media.

(Twitter Safety, 2017, para. 4)

Stage ten: CEO declared public accountability for progress.

- Management response

(March 1, 2018)

Twitter stated its new approach was to “improve the health of the conversation” on its platform as direct response to trolls. Jack Dorsey, co-founder and CEO of Twitter, stated, “We’re committing Twitter to help increase the collective health, openness, and civility of public conversation, and to hold ourselves publicly accountable towards progress” (Dorsey, 2018, @jack Twitter).
Stage eleven: Flagged accounts are muted.

- Feature response
(May 15, 2018)

Twitter introduced new features to fight malicious automation and spam. This feature muted tweets from trolls, making them less visible to users. This feature policy recognised that some trolling content did not violate any Twitter policy but did distort and detract from public conversation, particularly in communal areas like conversations and search. Twitter identified certain patterns of behaviour associated with troll-like accounts, including a user who signs up for multiple accounts at once, or users who repeatedly tag people that do not follow them in tweets. These ‘signals’ were identified by an algorithm and were tied to the behaviour of a user not the content of the tweets themselves. When identified, Twitter placed the account in ‘quarantine,’ where any of their interactions were selectively hidden from other users, unless specifically searched for by selecting the ‘show more replies’ button under a tweet. It is important to note this did not remove the comments, but the problematic content was obscured.

Stage twelve: Technology experts Smyte are introduced.

- Feature response
(June 21, 2018)

Twitter partnered with technology experts Smyte, which specialise in safety, spam and security issues, to combat trolling on the platform. Twitter released a statement, “Smyte has dealt with many unique issues facing online safety and believes in the same proactive approach that we're taking for Twitter: stopping abusive behaviour before it impacts anyone’s experience” (Twitter Safety, 2017, para. 3). Twitter further stated, "Smyte's products will help us address challenges in safety, spam and security more quickly and effectively. Their review tools and processes will be powerful additions to our own tools and technology that help us keep Twitter safe. We'll integrate this technology to strengthen our systems and operations“ (para. 4).
4.4.2 Thematic critical analysis

The next step was a thematic critical analysis of Twitter's responses to trolling. The responses were individually critiqued in chronological succession through which emerging themes within this timeline were integrated for discussion to provide a robust commentary for contextual analysis. The subsequent discussion begins by exploring themes that developed before the first documented Twitter response in order to provide the contextual framework for a judicious and vigorous evaluation. The study then provides a collective narrative of Twitter responses to trolling, each followed with a critique. Through the grounded theory approach, additional themes materialised, and these are discussed within the textual timeline of the progressive commentary. The thematic and critical analysis is predominantly qualitative; however, quantitative evaluation of time periods is also incorporated.

Initial slow response to trolling

The increasing presence of trolling on Twitter is well documented in chapter 4.1 findings (as described on p. 39); chapter 4.2 findings (as described on p. 64); and chapter 4.3 findings (as described on p. 82). Yet, no immediate remedy to increasing trolling was implemented by Twitter in response to calls for action. This study has identified key events and through quantitative analysis determined an approximate timeline of this initial response period. The study determined Twitter took 6.5 years initiate its first documented policy response to trolling after its launch on March 21, 2006. To analyse Twitter responses more specifically in the Australian news media, the study referenced the previously constructed Chronological timeline of the evolution of trolling in the Australian news media derived from chapter 4.1 data (see Figure 9 on p. 51). The study determined Twitter took three years and two months from the first documented The Sydney Morning Herald news article referencing trolling, published on July 15, 2009, before its first documented policy response to trolling on the September 14, 2012. An infographic to display these findings is presented in Figure 20 below.
Chapter 4.4 Twitter responses results and discussion

Figure 20

A timeline of Twitter’s first response to trolling

Note: The time lag of Twitter’s first response is illustrated by the addition of the bold black arrow lines. The snapshot displays the manifestation period and social problem period portion of the Chronological timeline of the evolution in the Australian news media derived from chapter 4.1 data in Figure 9 (on p. 51).

The study concedes it would be impossible to identify the first trolling incidence on the Twitter platform due to a variety of factors. Two primary complications include the vast amount of content on Twitter, and the subjective nature of trolling that results in discrepancies between what users deem to be harassment and what users considered to be a difference of opinions and vigorous debate. However, the first incident was determined to fall within the incubation period (as described on
The study can isolate Australia’s first trolling incident to within a three-year timeline: between Twitter's launch in 2006 and the first documented *The Sydney Morning Herald* news article referencing trolling, published on July 15, 2009 (Moses, 2009, para. 1). However, without this exact date, the study can still spotlight continued Twitter inaction. After the first referenced trolling articles in the Australian news media, numerous personal accounts by both Australian and international journalists were published in the *manifestation period* of this study (as described on p. 53). These echoed requests for Twitter to respond, yet these entreaties did not result in any publicly documented action. Twitter took a considerable amount of time to implement any policy or feature changes, although the increasing presence of trolls using its platform to find and abuse victims was indisputable if we take media coverage as the evidence.

**Beneficiary of trolling**

The study explored possible motives for the continued inaction and found evidence illustrating that Twitter was a beneficiary of trolling in the initial stages of its evolution. Some initial victims engaged in conversation with abusers as a defensive tactic (West, 2015), which further fuelled user interaction on the platform and increased the popularity of the platform at the time. Croteau (2019) suggested users have the tendency to overshare online which heightens the potential to reveal vulnerabilities open to criticism, and further proposed a large portion of content is posted for validation, whether sought consciously or unconsciously. Social theory goes some way to explain this desire for acceptance, be it in person or online. In a society “addicted to like” (para. 6), users are far more likely to respond to negativity over positive comments as a reactive process of our desire and fundamental need for acceptance. In addition to the briefly explored theory spotlighting user engagement with abusers, the increased frequency of trolling-related news articles published in the Australian news media during the *manifestation period* of the study (as described on p. 53) had little deterrent effect as the popularity of the platform continued to increase each year up until 2018 (Dean, 2022). The study, therefore, suggests that Twitter was a direct beneficiary of trolling in the initial stages of its evolution, and data suggests Twitter may have...
made a selective decision to delay its response to boost engagement (Phillips, 2015).

From the social problem period (as described on p. 55) that commenced in November 2011, the frequency of trolling in the Australian news media rapidly increased (as displayed in Figure 3 on p. 40) and news reporting exposed the severe negative impacts of trolling. The beneficiary position of Twitter that enabled trolls to flourish on their platform began to alter, as negative commentary permeated public discourse, with implications for the reputation of the service. The study identified widespread statements of users threatening to leave the platform including West (2017), Noyes (2017) and Vanian (2017). It was not until effects on Twitter's popularity and commercial viability were evident that any responsive action was initiated.

**Cumulative external pressure**

The cumulative pressure on Twitter began on August 30, 2011, when Sady Doyle’s #mencallmethings campaign went viral, flooding the platform with anti-trolling sentiments from users. Exactly one year later, after the attempted suicide of Charlotte Dawson on August 30, 2012, the #SayNoToBullying campaign also went viral, further escalating deeper anti-trolling sentiments and bringing these to the forefront of conversations beyond the platform itself, including to the Australian news media and the wider community. The pinnacle of Twitter’s externally mounting pressure in Australia emerged two weeks later on the September 11, 2012, when The Daily Telegraph published its own anti-trolling campaign #StopTheTrolls, which received global media traction, urging Twitter to act and reveal the identity of trolls to authorities for prosecution. This campaign provoked further calls for Twitter to act and is discussed in detail in chapter 4.6 (on p. 219).

On the same day:

- *The Daily Telegraph* published a further article that stated, “Twitter must be held accountable”.
- Acting Prime Minister Wayne Swan stated, “the #StopTheTrolls campaign is a very worthwhile initiative by The Daily Telegraph that the Australian Government fully supports” (Jones & Byrnes, 2012, para. 4).

*Chapter 4.4 Twitter responses results and discussion*
Communication Minister Stephen Conroy said, “Twitter may think they’re above Australian laws, but ultimately good corporate citizens do not behave this way. They should co-operate with law enforcement authorities to prosecute trolls who break the law” (Hildebrand & Matheson, 2012, para. 7). Senator Conroy further suggested, “Twitter should follow Facebook’s lead and establish a staff presence in Australia, so they are more receptive to community concerns” (para. 8).

It was not until the cumulative external pressure from journalists, campaigns, newsroom, and the government were directed towards Twitter that the organisation itself initiated a response to trolling. Twitter’s identified response with be discussed and critiqued below, from stage one to stage thirteen.

Stage one: Twitter’s first acknowledgement that trolling occurs on its platform

The first identified response to trolling by Twitter was categorised in this study as both a management response and a policy response. Immediately after direct calls to action from The Daily Telegraph, Acting Prime Minister Wayne Swan and Communication Minister Stephen Conroy, Twitter released its first response to trolling. The new 2012 policy was the first operational change by Twitter to combat trolling content on the social media platform. Twitter spokeswoman Rachel Bremer said, “we take this issue extremely seriously and have recently updated our abuse policy to make targeted abuse against our terms of service” (Hildebrand & Matheson, 2012, para. 3). However, in criticism of the statement, Hildebrand and Matheson (2012) suggest, “Bremer did not specify when the policy was updated or what specifically had changed” (para. 4). This study’s findings provide the documented data that revealed there was no policy in direct relation to abuse prior to Bremer’s statement. This policy ‘update’ was in fact a new policy that was not mentioned in the prior Twitter Rules. The first Twitter Rules were short and consisted of 568 words in total, with one sentence to describe the appropriate behaviours for each of the ten categories that this study has previously listed (see p. 90). One example was the thirteen-word Violence and Threat policy: “You may not publish or post direct, specific threats of violence against others” (Twitter, 2006, para. 1). The study acknowledges extreme
examples of trolling may have been covered within these initial rules, but that it was not until this abuse and harassment policy was announced on September 14, 2012, that Twitter addressed personalised attacks regarding gender, race, sexuality, appearance or political views. Trolling was recognised by Twitter to fall beyond the acceptable standard, and it should be reported for potential removal. Twitter’s (2012) new rule of engagement at the time clearly stated: “Technical abuse and user abuse is not tolerated on Twitter.com and will result in permanent suspension” (para. 1). This was a significant change from the previous guideline: “We do not actively monitor and will not censor user content” (Twitter, 2006, para. 1), which did not outline any repercussions for non-compliance. This evidence suggests that the “recent changes” Bremer was referring to were in response to the mounting pressure accruing from the #StopTheTrolls campaign that had been launched days earlier. This policy change enabled both users and Twitter itself to respond to trolls by removing their content and profiles from the platform.

Stage two: Twitter declares trolling as a continuing problem
The second response to trolling was categorised by the study as a management response and a policy response. Two years and one month passed between the Twitter’s first response to trolling on September 12, 2012, and the second response to trolling on August 11, 2014. This study identifies this time-period to advance from the social problem period to the social epidemic period. During the time-period between Twitter’s first and second response to trolling, many examples of trolling are documented in the news articles data set of this study, yet no further action was undertaken by Twitter. Emerging themes suggest Twitter was not proactive in the fight against trolls. Twitter was simply reactive after external pressures mounted to such a point that Twitter’s inaction began to tarnish the company’s standing. Only then were subsequent management statements made and new policies implemented. Left without external pressures, thematic analysis indicates Twitter may not have responded when it did, and it may have left trolls to flourish on its platform unmonitored.
Twitter's second response to trolling corresponded with a sharply increased frequency of articles in public discourse globally and locally, with discussions outlining not simply its abundance but also its associated problematic nature within the Australian news media and broader society in general. Twitter's second response to trolling corresponded with public outcry over Zelda Williams leaving the platform on August 13, 2014, after receiving horrific trolling abuse. Targeting online abuse of Zelda Williams was identified as a key event, further discussed below.

**Trolling of Zelda Williams**

Zelda Williams fell victim to abusive Twitter trolls after the death of her father, famous actor Robin Williams, on August 11, 2014. Williams reportedly received “one rape threat per minute” (Brown, 2014, para. 6) and announced she would leave the social networking platform for a “good long time” (para. 6). Although the trolling of Zelda Williams is globally recognised as a significant example of trolling affecting the evolution of trolling globally, the event was not covered in the Australian news media to the extent that it was evident in the *Chronological timeline of the evolution of trolling in the Australian news media derived from chapter 4.1 data* (see Figure 9 on p. 51). As the aim of this study was to illustrate the trolling evolution in the Australian news media, the dataset for the creation of Figure 9 focussed on profession (as a journalist) and location (in Australia) and publication (SMH, ABC or News.com) as the main inclusion criteria. Because this incident occurred in the United States of America, the study deemed the incident to fall outside of the Australian news media scope. However, as there was considerable relevant discourse (Dewey, 2014; Sparrow, 2014), it has been included in Figure 21 for context and further discussion and critique irrespective of its materialisation outside of Australian borders. From comparison, the Zelda Williams trolling has been added to a snapshot of Figure 9 and is displayed in Figure 21 below.
Figure 21

A snapshot of the trolling of Zelda Williams

Note: Zelda Williams’s trolling incident is highlighted in orange within the clustered examples marking at the peak of the social epidemic period. This snapshot displays a portion of the study’s Figure 9 Chronological timeline of the evolution in the Australian news media derived from chapter 4.1 data on (p. 51).

This particular trolling example received global widespread media coverage and set a turning point for public discourse expressing universal distress over Twitter’s inaction at the time. The study identified a number of triggering factors that provoked the widespread public outcry in support of Zelda Williams. Robin Williams (her father) was an admired actor, and the public mourning of his death
induced a sense of obligation to protect his daughter, as indicated in published online discourse. Williams was recognised for his extrovertedly happy demeanour, and it was additionally confronting when his private struggles with depression were revealed as the motives for suicide. This further fuelled the collective compulsion to protect his daughter. In a direct response to this public pressure, four days after the passing of her father, Twitter’s vice-president of trust and safety, Del Harvey, responded with the statement, “we will not tolerate abuse” (para. 3) and indicated Twitter would expand its policies regarding self-harm and private information and improve support for family members of deceased users. The timing of this announcement further illustrates its crisis communications response, and spotlights the causal relationship between the platform’s financial motives and its limited strategies to address trolling.

**Continued stage two: Twitter declared trolling as a continuing problem – critical analysis**

The study identified the notorious Gamergate Saga as an episode of extended, systematic, global trolling. While unrelated to the study’s focus on the trolling of Australian journalists, it was instigated on the same day as Twitter’s second response detailed above and therefore was of relevance. The Gamergate Saga was included in the study’s *Chronological timeline of the evolution of trolling in the Australian news media derived from chapter 4.1 data* (see Figure 9 on p. 51) as a universally recognised example of severe, targeted abuse that defined the evolution of trolling on a global scale. Although the Gamergate Saga itself began in the United States of America, a significant portion of critical research about the incident has been conducted by Australian academic Dr Emma Jane, an Associate Professor at University of New South Wales. Jane’s (2015) extensive investigations and analysis of the Gamergate Saga has been both published and cited by others in journal articles, news reports, and in academic discussions on the topic of trolling globally. For this reason, this study selectively included the Gamergate Saga in the timeline although its materialisation fell outside of Australian borders. On August 15, 2014, troll Eron Gjoni published his first diatribe against ex-girlfriend Zoe Quinn on multiple websites, including Something Awful and Penny Arcade. After receiving an immediate ban, Gjoni
continued the onslaught of abuse through anonymous message boards including 4chan, Reddit and Twitter. However, the most prolific vitriolic harassment emerged on the Twitter platform. It is important to note that the Gamergate Saga did not have a direct impact on Twitter’s response strategy, as Gjoni published his first trolling attack on the same day that Twitter announced its second policy update. However, it does spotlight the peak of the trolling era in the Australian news media identified by this study as the *social epidemic period* (described on p. 61). This study therefore suggests the first two response strategies Twitter had implemented in 2012 and 2014 were insufficient to adequately address the surge of trolling on its platform, as examples of sadistic trolling persisted and flourished across the platform, including the Gamergate Saga. According to academic and criminologist from the University of Western Sydney, Michael Salter (2010) Twitter was unable to protect users from their abusers and facilitated the snowballing trajectory of trolling, through functionality flaws that resulted from unchanged features and futile safeguarding policies. These functionality flaws enabled trolls to overwhelm users' ability to individually block the large number of fake accounts used to distribute malicious messages. This study’s findings echo Salter who, in his published 2019 research *Crime, Justice and Social Media*, labelled Twitter’s architectural design as "conducive to abuse and exploitation" (p. 19). Salter also advocates that delays in Twitter’s trolling response were financially motivated as this study also suggests. Salter explains:

Crucial to this discussion is the status of social media as a network of corporate platforms that profit from the commodification of user data. Communication of social media is induced by software architecture that actively encourages the publication and circulation of private, emotive, or provocative material (such as trolling) that drives market share and revenue. It is within the tension between communication and commodification that online abuse takes shape and meaning, and exerts its impacts on users and public debate.

(Salter, 2019, p. 18)

This study suggests that Twitter had not actively engaged with the growing concerns for calls to action from numerous stakeholders, including its users, journalists, academics, news organisations, governments, and the broader community. As the Gamergate Saga transcended its original victim, Zoe Quinn, to
encompass other female game developers, such as Brianna Wu, demands for Twitter to act were loud (Stuart, 2014). This was reported across the Australian news media (Noyes, 2014; McGrath, 2014), yet no further action resulted. In an interview published by *The Guardian*, Brianna Wu suggested Twitter was responsible for facilitating harassment and challenged the platform to improve its responsiveness to complaints (Stuart, 2014). This study suggests further Twitter’s policy and feature updates were in reaction to the growing dissatisfaction of users, leading to a decrease in usership and subsequent decreasing revenue. As a revenue-led company, Twitter was reactive in its response to trolling.

**Legal exemption**

The study acknowledges Twitter’s legal exemption from media regulation in the US as a contributing factor to the platform’s delayed response to trolling. Twitter has no legal obligation to police malicious content such as harassment and threats, as protection by US law states that social media platforms assume no liability for content posted by third-parties of their service under Section 230 of the Communications Decency Act (1996). Likewise, Australian laws do not provide a universal solution to suggestions of platform liability, and the debate in 2019 surrounding defamation law reform continues. On November 20, 2019, Australia’s Attorney General Christian Porter announced plans to make platforms such as Twitter and Facebook liable for the content posted by third parties as part of a wide range of planned defamation law reform (Taylor, 2020, para. 8). The first tranche of legislation was due to be introduced in June of 2020, but the Covid-19 pandemic delayed the government’s action, and no resolution was yet known at the conclusion of 2021. In relation to such legislation, the study addresses legal responses as a standalone stakeholder, including interviews with key figures in media legal ethics, at length later in the discussion chapter (see section 4.7 on p. 265).

**Stage three: The reporting method for abusive content is streamlined**

Twitter announced a collaboration with Women, Action and the Media (WAM, 2014) on November 6, 2014, that encouraged users to report harassment on an external website for severity assessment before selected escalation back to...
Twitter. This study excluded this collaboration from Twitter’s documented
trolling responses for two reasons. The New York non-for-profit group was not
considered to fall under the Australian scope and, unlike Zelda Williams, did not
have considerable relevant discourse in the content analysis to validate inclusion
in this study. Additionally, the external reporting function that was provided gave
negligible additional service over Twitter’s direct reporting function that was
already provided. Attempts to outsource its content regulation did nothing to
address the intrenched architectural issues of the platform itself, it just created an
additional inbox for complaints to be submitted to. The collaboration is noted,
however, but it was not included it in the documented list of Twitter responses to
trolling.

The third response to trolling, (categorised by this study as a feature response),
occurred in 2014 when Twitter received widespread criticism for its poor
handling of trolling on the service. Technology publication The Verge described
the problem as Twitter’s “weak response to violent threats” (Newton, 2014, para.
1). Subsequently, on December 2, 2014, Twitter announced four feature changes
for reporting of abuse on the platform, which circulated 500 million tweets every
day at the time (Twitter, 2014, para. 2). The first improvement was to streamline
the harassment reporting process, which had previously required a nine-part
questionnaire, as displayed in Figure 22 on the following page.
Figure 22

Twitter’s nine-part questionnaire for reporting abuse in 2014

This image was produced by Twitter in 2014 on their blog (Twitter, 2014).

Although the process still required complaints to be reviewed manually, the condensed method enabled Twitter to identify and remove trolling content and perpetrators from the platform more efficiently (Twitter, 2015). In addition to the shortened reporting process, users who observed abuse targeted towards others
but who were not directly involved, could flag trolling content. Improved algorithms considered behavioural signals that now included reports from bystanders and used those signals to sort and prioritise the severity of threats for faster response. For example, if 100 users all reported the same tweet, the algorithm would escalate an expedited response. Twitter’s emerging challenge was to ensure users were not “gaming the system” by reporting as abusive the tweets they simply disagreed with (Newton, 2014, para 5). Twitter also introduced a ‘blocked accounts’ page displayed in Figure 23 below. This feature allowed users to view and edit accounts they had previously blocked providing control over specific content viewed by users.

**Figure 23**

*Twitter’s block feature introduced in 2014*

![Twitter’s block feature](image)

This image was produced by Twitter in 2014 on their blog (Twitter, 2014a).

In addition, Twitter prevented blocked users from viewing the profiles of the people who blocked them in an attempt to protect the privacy of its users. In a corresponding statement, Twitter said “we are nowhere near being done making changes in this area” (Twitter, 2014, para. 5), admitting further changes were still required to adequately address its continued trolling problem.
Stage four: CEO declares responsibility for trolling on the service

The fourth response to trolling was categorised by the study as a management response. By February 2, 2015, the Gamergate Saga had escalated to a global scale, and the evolution of trolling in the Australian news media had surpassed the halfway point of the study's identified social epidemic period (described on p. 61). Trolling examples were evident on this particular microblogging platform, yet no viable Twitter responses to eliminate its presence had been successful.

In direct response to Lindy West’s article, What happened when I confronted my troll cruellest troll, which was published earlier on this same day, February 2, 2015, Twitter’s CEO Dick Costello admitted on his Twitter account the platform had failed trolling victims (as noted in full on p. 107). Key remarks of the 127-word admission included, “we suck at dealing with abuse”, “I’m frankly ashamed of how poorly we’ve dealt with this issue”, and “it’s nobody else's fault but mine, and it’s embarrassing” (Costello, 2015, @dickc Twitter). Although published by the UK-based news organisation The Guardian, West’s (2015) article was the inspiration for this study and therefore has been included in the study’s Chronological timeline of the evolution of trolling in the Australian news media derived from chapter 4.1 data (see Figure 9 on p. 51) for comparison. This notable incident sparked academic criticism and debate amongst journalists about trolling as an acknowledged part of the Australian media workplace environment at the time (Tiku & Newton, 2015). West inspired Australian journalist Noyes (2017) to speak out about the impacts of daily rape threats and death threats, which are an acknowledged although uncomfortable and toxic element of the job (as discussed in section 2.5.1 on p. 24). Although Costello’s “we suck” statement did not coincide with any feature or policy changes, it was the first public statement in which the platform took responsibility for trolling. The following day, February 3, 2015, two subsequent statements further declared Costello’s “personal responsibility” for the prevalence of trolling on Twitter; however, no functional changes were announced. Three days later, on February 6, 2015, Twitter released a report that labelled trolling as a serious threat to its popularity and subsequent profitability and reported a decline in active users. Twitter’s average monthly users fell by 4 million in the final three months of 2014, down to 288 million users (Twitter,
The sudden quick succession of Costello’s responses preceding the damning report may reflect his prior knowledge of the report and a financial motivation for Costello’s statements. Twitter’s sudden response to trolling, therefore, seemed to not solely emerge from inherent or external ethical pressures, but to be driven by financial losses from decreased usership and subsequent reduced profit margins. Once online abuse began to affect the bottom line of the company, its efforts to resolve it were visibly accelerated. This study identified that amongst Twitter’s earliest responses to trolling, the first, second and fourth responses appear to be clearly financially motivated.

**Stage five: a phone number verification is required for all users**

The fifth response to trolling was categorised by the study as a feature response. Newly declared financial declines drove a swift response to trolling that the previous six years and eight months of ethical pressures could not alone achieve. After declaring the financial impact of trolling on February 6, 2015, Twitter took just 35 days to respond on March 13, 2015, to combat trolling. In comparison, Twitter took nearly seven years from the first documented Australian trolling example in the *The Sydney Morning Herald* on July 15, 2009, before this feature response to trolling changed on March 13, 2015.

Twitter now required a phone number verification for all new accounts. An email address was all that was previously required, and new accounts could be obtained with no verification. In support of Twitter’s response, IT publication *The Verge* explained “email addresses are relatively easy to obtain, but phone numbers are harder. By checking the phone numbers against a list of banned users, Twitter could be able to keep more abusers and harassers from creating accounts” (Newton, 2015, para. 1). In cases where a user was found to be harassing others, their account could be temporarily banned until a phone number was provided and checked against a list of known trolls to decide whether a full ban was necessary (Russell, 201, para. 4). The new verification method was specifically introduced to combat fake accounts used by trolls, a problem Brianna Wu had flagged and publicly urged Twitter to address in the previous year (Stuart, 2014). This was the first response that required accountability of users; however, critics
at the time suggested users could simply purchase a pre-paid sim card or use an app such as Burner, which provides a temporary phone number free of charge as a loophole (Higgins, 2015). In support, technology critics TechCrunch agreed and labelled the response “not a perfect solution, but a positive step forward” (para. 4).

**Stage six: the quality filter feature is implemented**

Twitter's sixth response to trolling was categorised as a feature response and came in quick succession, ten days after the previous feature change announcement of stage six. On March 23, 2015, Twitter implemented the Quality Filter, aimed to remove all tweets that contain threats, offensive or abusive language, duplicate content, or are sent from suspicious accounts from users' notifications timelines (Welch, 2015, para. 1).

Technology critics at the time suggested this was “not a radical new tool” (Shu, 2015, para. 4), but rather an extension of another feature for verified users called “tailored filtering” (para. 4), which allowed users to select an edited version of their timeline notifications based on certain factors, with the most influential being interactivity frequency. Content from accounts with most interactions were displayed on the platform, and accounts with no engagement were shadow banned. The new quality filter extended this functionality to all notification options, including the expletives and vulgar language commonly featured in trolling. From a strategic standpoint, it was logical for Twitter to make this feature available for verified users, primarily because these accounts had the most followers and garnered the highest proportion of trolls (Shu, 2015, para. 5). However, genuine non-verified users also had to cope with bullying on the platform and were not protected by this feature.

On July 24, 2019, Twitter announced, “we're removing Quality Filter from search settings because people told us the filter was confusing and they didn’t want it” (para. 1). Tech critics suggested the feature may not have been “confusing”, but rather that the functionality of its algorithm had difficulty in practical implementation, as the company never offered any details on the algorithms used.
to help determine what was considered a low-quality tweet (Miller, 2019, para. 3). The filter may have prioritised accounts that users were uninterested in, reducing the platform popularity amongst users that thrives with provocative content. The study proposed Twitter should have reworked its algorithm in order to achieve its initial objectives rather than removing the feature.

Stage seven: new accounts filtered for previously flagged content
Twitter’s seventh response to trolling was both a policy and feature response aimed to reduce hateful conduct and abusive behaviour by revising a former policy on April 22, 2015, expanding its description to further state users could not “direct specific threats of violence against others or promot[ing] violence against others.” The amendment meant any user who shared another user’s threat was now infringing the Twitter policy, which became a problem for feminists who shared abuse to shame the abuser. This updated policy aimed to eliminate the loophole used by trolls to circulate other users’ threatening comments, which allowed accounts to spread trolling content without personal liability or repercussions. The implementation of this policy also aimed to reduce the potential influence of mob mentality identified in interviews with stakeholders and analysis of online commentary, that created bonds and a sense of camaraderie between trolls, as well as offering a sense of support between abusers. Mob mentality, also known as pack or herd mentality, was identified through thematic analysis as a critical underlying theme for further analysis and discussion below.

Facilitating online mob mentality
Biological evidence denotes humans as social beings, with the lateral prefrontal cortex of the brain often enforcing compliance with social norms (Yin et al, 2017). This compliance is an important part of evolutionary psychology, as complex social structures are the foundations of survival (Ruff, Ugazio & Fehr, 2013). However, this fundamental conformity can lead to the acceptance of, or participation in, emerging negative behavioural patterns such as trolling.

Social influence literature has shown that people choose to conform, because they explicitly desire to adhere to the expectations of their group, in a process of
“normative influence” (Deutsch & Gerard, 1955). This decision is often an unconscious, automatic process resulting in behavioural mimicry. Imitation is traditionally characterised by personal interactions; however, the online domain has facilitated the means for virtual transgression. In a society “addicted to likes” (Croteau, 2019, para. 6), the desire to be accepted can impact the ability of users to be objective in the face of harassment and may contribute to heightened tolerance of fear of rejection within the social dynamics that online forums provide. In a 2014 study of social defaults, Hu et al concluded, “just as observing others’ behaviour can induce behavioural mimicry, observing others’ choices can induce choice mimicry and may cause these choices to become default options” (p. 747). This description suggests social theory concepts materialise online and have the ability to shape the behaviour of platform participants. Unmonitored, this influence has the potential to accelerate the manifestation of a social problem, masked by a shield of perceived tolerance.

The social context of online platforms such as Twitter can exogenously create vitriolic environments leveraged by online abusers. The unconscious nature of this process may be a contributing factor to its lack of recognition in current strategies to address trolling and is a suggested area for further investigative research in practical application. Understanding the influence social default has on behaviour, both on and offline, consciously and unconsciously, is critical to combat the threatening and degrading rhetoric disseminated on Twitter.

**Continued stage seven: new accounts filtered for previously flagged content**

In addition, Twitter announced a feature to help identify abusive content and limit its potential reach by analysing the age of Twitter accounts and comparing the similarity of content between previously reported threats. This feature was designed to stop trolls immediately setting up another account when they have been previously reported and blocked. These fake accounts were recognised as troll minefields years prior (Seddon, 2014). Despite this, Twitter delayed the implementation of any combative strategy to the known problem until April 22, 2015, nine years after Twitter’s initial launch, and eight years and nine months
after the first documented trolling article in the Australian news published in *The Sydney Morning Herald* on July 15, 2009. This study again highlights and criticises Twitter’s continued delayed response to effectively keep trolls and their vitriolic behaviour off the service. This new feature allowed Twitter to freeze an account until the user deleted the flagged tweet. This study criticises Twitter’s soft resolution approach to ‘freeze’ accounts with flagged content, in preference to an immediate deletion. However, the study further acknowledges the wicked aspect of content regulation makes it impossible to completely ban individuals if they have access to new account creation via phone spoofing and disposable email addresses. The study suggests the continued delays of responses indicate Twitter is a benefactor in the circulation of disruptive content to increase circulation and use of the service.

**Stage eight: the Trust & Safety Council is introduced**

The eighth response to trolling was categorised as a management response. Seven of the thirteen identified stages developed within the *social epidemic period* of the study (described on p. 61): from Twitter’s second response to trolling on August 11, 2014, marking the beginning of *stage two*, through to the eighth response on February 10, 2016, marking *stage eight*. As trolling incidents flooded the platform, Twitter announced the implementation of a Trust & Safety Council, as the *saturation period* approached (described on p. 62). Introduced to monitor harassment and cyber bullying, the Council was initially made up of 40 organisations from around the world, four of which were Australian-based (Ockenden, 2016). By 2020, the council comprised of 48 organisations, five of which were Australian-based. These Australian organisations were: The Alannah and Madeline Foundation; Beyond Blue; Bravehearts; Project Rockit; and ReachOut Australia but in 2019 Council members said the group was ignored (Matsakis, 2019).

**Stage nine: accounts that promote violence are prohibited**

Twitter’s ninth response to trolling was a policy response that developed at the end of the *saturation period* (as described on p. 62) when trolling examples in the Australian news media were most frequent. On November 18, 2017, Twitter
updated its policies with a new rule against violence and physical harm. The new policies prohibited specific threats of violence and abusive behaviour that harassed, intimidated, or used fear to silence another user on the service. In addition to offensive text, hateful imagery was also recognised as hostile and outlawed from the service (Twitter Safety, 2017). All malicious content reported to Twitter in breach of the new policies would now result in the author’s permanent suspension.

The harder stance was a significant jump from *stage seven*, where troll accounts were reprimanded with a “freeze” until the reported content was removed, and the user could continue using the service. This was the first Twitter response to directly acknowledge targeted abuse toward minority groups and labelled the subcategories of “race, religion, disability, sexual orientation and ethnicity” (Twitter Safety, 2017, para. 3) as areas of particular concern, identified and supported by the findings in chapter 4.2 (on p. 64). Although numerous examples of racist trolling (Blackwell, 2015), religious trolling (Pearson, 2012), disability-based trolling (Bullen, 2016), sexuality-based trolling (West, 2015; Ford, 2016) and ethnicity-based trolling (Phillips, 2015) were evident years prior and provided clear evidence of targeted abuse towards marginal groups on the platform, it was over a year into the *saturation period* before Twitter finally made the acknowledgement and initiated the policies in responsive action.

**Stage ten: CEO declares public accountability for progress**

Twitter’s tenth response to trolling and was a management response. On March 1, 2018, Twitter stated their new approach was to “improve the health of the conversation” on its platform as a direct management response to trolls. Jack Dorsey, co-founder and CEO of Twitter, stated, “we’re committing Twitter to help increase the collective health, openness, and civility of public conversation, and to hold ourselves publicly accountable towards progress” (Dorsey, 2018, @jack Twitter). With this statement, Twitter declared public accountability for the progress towards identifying and removing trolls from the platform.
This response was the first recorded beyond this study’s thematic content analysis displayed in Figure 9: Chronological timeline of the evolution of trolling in the Australian news media derived from chapter 4.1 data (on p. 51). Although Twitter had previously declared responsibility for trolling on the service in on February 2, 2015 (as identified in stage four), three years and one month later the platform extended their declaration from responsibility to accountability for abuse on their service. However, the profound statement alone did not offer any support for victims. The statement did not provide an action plan and did not define any measurable ways to access the platform’s progress. In short, the statement was one of hope rather than an accountable progression method.

**Stage eleven: flagged accounts are muted**

Twitter’s eleventh response to trolling was a feature response. On May 15, 2018, Twitter introduced new techniques to fight malicious automation and spam by muting the content tweeted by trolls on users’ accounts. This function did not remove trolls from the platform, though it made them less visible to users. Twitter identified certain patterns of behaviour associated with troll-like accounts, including users who sign up for multiple accounts at once, or users that repeatedly tag people that do not follow them in tweets. Identifying emerging behavioural characteristics was developed by Twitter as a key tactic to isolate trolling accounts, placing flagged accounts into ‘quarantine’. If a user specifically selected the ‘show more replies’ button under a tweet, all comments would appear including those of trolls.

In critique, the study suggests the muting function of flagged accounts was not a standalone feature response, but an extension of two previously identified stages. The first was an extension of the stage seven: new accounts filtered for previously flagged content. Literature surrounding the muting feature response named two key behavioural patterns the algorithm scanned for. The first behavioural pattern the algorithm scanned for was users who sign up for multiple accounts at once. The second behavioural pattern the algorithm scanned for was people who repeatedly tag people that do not follow them in tweets. This was a new function not considered by any of the prior stages.
The second was an extension of stage six, the ‘quality filter’ feature, that aimed to remove all tweets from users’ notifications, timelines that contained threats, offensive or abuse language, duplicate content, or were sent from suspicious accounts. However, this feature did not remove all trolling content, as some content, although offensive, did not breach any of Twitter’s policy and this is why it was not removed. From cracks in the ‘quality filter’, the ‘muting function’ followed, where content that was not deemed offensive enough to be removed from the platform but was still problematic, was muted. The study proposes the argument that if content was deemed problematic enough to be flagged and hidden, why was it not deemed offensive enough for removal by Twitter in the first place? Twitter’s choice to mute rather than remove trolls suggests financial benefit was again a motivation for inaction, as interactivity from active debate about trolling on the platform led to increased usership and subsequent increased profit margins.

Stage twelve: technology experts Smyte are introduced

Twitter’s twelfth identified response to trolling was categorised by the study as a management response and a feature response. On June 21, 2018, Twitter announced its partnership with technology experts Smyte, who specialise in safety, spam, and security issues, to combat trolling on the platform. Although the statement was made by Twitter’s management team and did not disclose any exact feature changes, the study identified the operational and technical elements provided by Smyte (not detailed in this study) would additionally classify the change as a feature response. In the same statement, Twitter declared the health of the conversation remained their top priority. However, the continued presence of trolling on the platform, coupled with the minimal implemented response strategies documented in this study, would challenge that claim. Twitter’s partnering with an external company may suggest it was struggling to effectively control the escalating trolling problem and that its internal structure could not cope with the onslaught of trolling. This further suggests that previous response attempts by the company were insufficient. Having outsourced its own regulation may suggest Twitter was ill-equipped to have any effective impact on trolling and may have contributed to its continued delayed reactions. In this study, both the
viability and ethics of continuing to operate a company unable to control such a persistent snowballing problem is brought into question.

Twitter response excluded from data set
Twitter further stated an additional response to trolling on June 26, 2018, to combat malicious automation and spam; however, this study does not consider the “four new measures” to be new features. Twitter’s announcement outlined previous stages and provided no additional policy or feature changes. The announcements are described as follows.

1. The first measure was to “reduce the visibility of suspicious accounts in Tweet and account metrics”, reinstating stage eleven’s flagged accounts as muted section.

2. The second measure was to “improve the sign-up process, with new accounts being required to confirm either an email address or phone number when they sign up”, reinstating stage five’s phone number verification requirement for all users.

3. The third measure stated, “auditing existing accounts for signs of automated sign-up” and restating stage seven’s filtering of new accounts for previously flagged content section.

4. The fourth measure focussed on the “expansion of our malicious behaviour detection systems to include automated detection of suspicious account activity,” including automatically flagging Twitter accounts that tweet an exceptionally high-volume of the same hashtag, or the same @username without a reply from the account mentioned, restating stage eleven’s muted flagged accounts. In this section, Twitter stated two behavioural patterns that the algorithm scanned for. The second behavioural pattern was people who repeatedly tag people who do not follow them in tweets. This proposed “new measure” was therefore deemed not to be new.

With the frequency of trolling continuing to increase at this time (Dean, 2022), Twitter’s inability to implement any new policy or feature changes indicate that the platform is incapable of delivering a universal comprehensive prevention response strategy for victims, or that it is unwilling to.
Effectiveness of Twitter responses

The nature of grounded theory research is difficult because social problems are messy. This study has gone some way towards identifying the evolution of responses over time; however, further research would be needed to establish which responses are most effective. The study identified that Twitter’s responses were designed with two objectives: firstly, to protect users from trolls, and secondly, more often to protect Twitter from the bad press, corporate damage and loss to share prices associated with hosting trolling. Twitter’s initial three responses deflected responsibility and avoided any significant action. Nine years after its launch, Twitter finally declared partial responsibility for trolling, and its ensuring proceeding responses evolved from self-preservation, to more actively protecting its users. However, response strategies were not considered by this study to be proactive, but to be reactive to a financially undesirable event. Reflecting on the practical application of Twitter’s responses to trolling, the study offers the following findings:

- Management responses have the least potential impact to safeguard users from trolls.
- Policy responses have a moderate potential impact against trolls by prohibiting their behaviour.
- Feature changes have the greatest potential impact as a combative strategy to eliminate trolling on the platform.
- Future research may be most effective if directed towards how to best to implement feature strategies on social media and measure their actual impact.

A robust and comprehensive evaluation of the effectiveness of Twitter responses to trolling is beyond the scope of this study. Further research would be required for broader conclusions about effective practice to be drawn. However, by reflecting on the findings the study can provide preliminary suggestions, as follows:
1. **To protect the Twitter platform from trolls**, the study’s preliminary findings suggest the following features may be the most influential existing strategies, and are considered of equal importance:
   - *Stage five’s* feature that requires a phone number for verification.
   - *Stage seven’s* feature that filters new accounts for previously flagged content.

2. **To protect Twitter users from trolls** the study’s preliminary findings suggests two complementary features in a synthesised approach may have the greatest potential impact. In a two-step process, the first suggested step is the removal of trolling content that contains abuse language, duplicate content, or that is sent from suspicious accounts achieved when users implement:
   - Stage six’s quality filter feature.
   The second consecutive suggested step is to hide any trolling content that was not removed by the previous step, achieved when users implement:
   - Stage eleven’s feature that mutes flagged accounts.

3. **To protect Twitter users from trolls**, the study’s preliminary findings suggest the response with the greatest potential to combat trolling with a standalone strategy is:
   - Stage six’s quality filter feature.

This chapter has outlined Twitter’s ten types of responses to trolling, providing clarity about an important stakeholder in the trolling paradigm. While many other companies offer training for journalists in the safe and effective use of Twitter (Duggan, 2017; Committee to Protect Journalists, 2018; International Federation of Journalists, 2019), this study found minimal evidence of action on Twitter’s behalf between 2009 to 2019 to protect journalists from trolls. The next chapter will investigate the other leading social platform Facebook and its responses. These analyses provide a grounded view in line with the systems model that suggests social problems require actions on multiple levels. In this analysis the social platforms sit arguably between governance and support levels of the
trolling paradigm with actions by fellow users falling under support, and regulatory actions by the social media companies functioning as governance.
Chapter 4.5

Facebook Responses

Results and Discussion

Thematic content analysis of Facebook responses

*Question E - What supportive strategies, operational modifications and corporate reform were implemented by Facebook in response to trolling?*

As previous chapters have demonstrated, the prevalence of online abuse and harassment and its disproportionate adverse impact on women has been evident since the internet's popularisation in the 1990s. Facebook is one of the leading social media platforms for driving traffic to news websites in Australia, with 33 percent of Australians using Facebook for general news in 2021, down from 46 percent in the early days of this study in 2016 (Park et al, 2021). This chapter analyses Facebook’s publicly stated intentions to manage trolling.

The research design strategy that addresses the research question: *What supportive strategies, operational modifications and corporate reform were implemented by Facebook in response to trolling?* applies mixed methods of thematic content analysis of archival evidence. The study used collected data from some of the articles collected for the analysis presented in chapter 4.1 (see p. 30) as a starting point. Of the 528 articles in the thematic content analysis, 31 mentioned Facebook’s response to trolling on its platform in some capacity and were deemed relevant for inclusion. Of these, 14 further discussed or made comment on aspects of particular response strategies. To obtain further data, Facebook’s website was explored, and in May 2020, relevant information was retrieved from three sections: About Us, Safety Centre, and Help Centre. On each of these pages the search function was used to investigate the terms: “troll” “bully”
and “abuse”. This led to a further 25 relevant articles that were included for analysis, bringing the total to 56. Although Google was also used as a search tool with the search inquiry: “troll” + “Facebook” + “response”. In May 2020, this retrieved 3,520,000 results, on closer analysis of the first ten pages totalling 100 articles, no new response strategies were found that had not already been identified in the previous methods of this chapter. Thematic analysis of the 56 articles grouped Facebook responses into five categories: support strategies, feature strategies, functional strategies, research strategies and corporate approach strategies, which are independently explored in the sections that follow.

Determining data for inclusion

The study first determined the parameters governing information that was considered to be trolling-related and, therefore, included in the data set. It was noted that Facebook often used the terminology ‘cyber-bullying’ in preference to ‘trolling’ in the material on its website. Therefore, the research analysed the content of each piece was analysed to determine which of the cyber-bullying data should be included in the data set rather than filtering for the exact terminology used by Facebook. The literature review of this study evaluated the various definitions of trolling cited in the literature and determined that trolling is best defined by Johnston (2014) who described it as “a deliberate attempt to upset or anger someone online” (para. 4). Johnson’s description was selected as the key definition of the study to determine which pieces of content that Facebook categorised as cyber-bullying would be classified by the study as trolling. Through this deductive method, qualitative and quantitative data was identified for thematic analysis, for discussion and critique.
Facebook responses strategy outline
The study applied thematic analysis to responses to trolling retrieved from archival evidence on Facebook, revealing five categories listed below. For a comprehensive analysis, individual response strategies were collectively analysed within their respective categories, providing robust data for in-depth exploration, discussion, and critique.

Thematic analysis of the 56 articles retrieved as archival evidence grouped Facebook responses into five categories which are independently explored in the sections that follow.

**Facebook support strategies**  
Page 142

**Facebook feature strategies**  
Page 161

**Facebook functional strategies**  
Page 174  
(automated software to the backend process)

**Facebook research strategies**  
Page 196

**Facebook corporate approach strategies**  
Page 203
4.5.1 Facebook support strategies

Facebook support strategies in response to trolling were explored first. The study identified six support strategies that form the qualitative data results of this section. Itemised in chronological order below, the list outlines the resources implemented by Facebook that Australian journalists can use when trolled.

1. **Facebook Safety Advisory Board**
   - Initiated on December 6, 2009

2. **Facebook Safety Centre**
   - Initiated on April 13, 2010

As an extension of the established Safety Centre, Facebook integrated the additional two resources.

3. **Facebook Parents’ Portal**
   - Initiated on December 13, 2016

4. **Facebook Youth Portal**
   - Initiated on May 15, 2018

Despite the introduction of the Parents’ Portal and Youth Portal, these resources are not particularly relevant to Australian journalists seeking to access information in response to being trolled as part of their work. The portals are cited here to illustrate the adoption a multi-levelled systems approach by Facebook.

5. **Be Bold Stop Bullying Campaign**
   - Initiated on November 2, 2012

6. **Facebook Bullying Prevention Hub**
   - Initiated on November 6, 2013
Emerging themes within each of the six identified support strategies are individually analysed, discussed and critiqued in the following sections.

**Facebook Safety Advisory Board support strategy background**

The Facebook Safety Advisory Board was the first support strategy implemented in response to trolling. To provide context for analysis, the study provided a brief background and summarised its role below from descriptions on Facebook’s website: https://www.facebook.com/help. Launched on December 6, 2009, the Facebook Safety Advisory board was comprised of leading internet safety organisations that provided expertise, perspective and insights to inform, guide and shape the approach to online safety on the platform (Facebook, 2009). Established to advise and implement best-practices on the service, the board suggested resources, tips, and tools for users to implement in response to trolling (Facebook, 2020a). These suggestions were refined and regularly updated. However, it is unknown if Australian journalists viewed or used them.

**Facebook Safety Advisory Board support strategy quantitative data**

The study analysed which countries were represented on the Facebook Safety Advisory Board. The first panel was comprised of five organisations, all of which were from western cultures:

- Four were based in the United States of America.
- One was based in England.

By January 1, 2020, after eleven years of operation, the number of organisations on the board had risen from five to eleven. Although representative countries had diversified over time, they were still predominantly western, representing 72% of the panel (eight of the eleven):

- One was based in Australia named Project Rockit. However, this organisation aimed content towards children and youth educators which was not relevant to working journalists (Project Rockit, 2020).
- One was based in New Zealand.
- Three were based in the United States of America.
- Two were based in England.
• One was based in Europe (with three office locations in France, Luxemborg and Portugal).
• One was based in Taiwan.
• One was based in Brazil.
• One was based in India.

While Facebook claimed in 2020 that the Facebook Advisory Board consisted of “organisations from around the world” (2020a, para. 1), 193 countries are recognised by the United Nations, yet only 10 countries are represented on the board. An analysis of global usage of Facebook (Statista, 2021) reveals that seven of the top ten countries with the highest number of Facebook users were excluded from representation on the board. The findings revealed that Facebook’s claim was misleading and did not accurately reflect the true distribution of users around the world.

This disproportionate representation was further explored. For comparative analysis, Australian Facebook usage data was identified and calculated as a representative portion:

- Global Facebook users 100% 2.603 billion
- Australian Facebook users 0.79% 11.23 million

The findings revealed that Australians comprised 0.79% of Facebook users, yet comparatively represented 9% (1 of the 11) of representative organisations on the Facebook Advisory Board. The findings support the suggestion of a western bias.

In contextual consideration, India was identified as the country with the highest number of Facebook users, and their representative portion was calculated for further comparative analysis:

1. Indian Facebook users 17.2% 346.2 million

The findings revealed that India as a standalone country comprised 17.2% of total Facebook users yet represented 9% (one out of eleven) of the Facebook Advisory Board organisations.
Board, equal to Australia. The findings identified a biased skew towards the inclusion of affluent nations on the Facebook Advisory Board. Only three of the eleven organisations (27%) specifically the ones from India, Taiwan and Brazil, represented developing nations. The study suggests this substantial skew does not accurately reflect the global distribution of Facebook users, and subsequent support strategies provided by the board may not proportionally represent the most effective responses to trolling for non-westernised regions.

**Facebook Safety Advisory Board support strategy thematic analysis**

Through thematic analysis of quantitative data results about the Facebook Safety Advisory Board from 2009 to 2020, four themes emerged including western bias, internet censorship, discriminative leverage, and affluence bias that are discussed and critiqued in the following sections.

**Western bias**

With five of the initial organisations on the Facebook Advisory Board based in the United States (80%), a western bias was established within the foundational processes that govern the site. Considering the company was founded in Cambridge, Massachusetts in January 2004, it is a logical that the Advisory Board was first established in the US. However, as usage of Facebook spread globally, the number and location of organisations represented on the board diversified but did not expand in proportion to the service's global distribution. On critical the examination of data and subsequent statistics, the study suggests that the appointed Facebook Advisory Board in 2020 did not accurately reflect a representative sample of global users of the service. This indicated that Facebook’s claim to represent organisations from “around the world” as misleading. This allows for trolling of journalists based on race to be unchecked.

**Internet censorship**

Although the scope of this study does not permit a comprehensive analysis of global internet censorship, it cannot critique the disproportional representation of the board, and Australia’s subsequent proportional representation within it, without first considering which countries have access to the platform. An
An infographic from a 2019 study by Freedom House illustrates the comparative variances of the identified accessibility theme for contextual analysis. Figure 24 displayed below depicts contrasts within global internet censorship, and spotlights jurisdictions of particular authoritative control.

**Figure 24**

*The state of internet freedom around the world in 2019*

The map indicates the global distribution of internet freedom identified by the study, evaluated on the basis of three categorical considerations: obstacles to internet access, limitations of online content, and violations of user rights. The report highlights China as the worst abuser of internet freedom from 2016 to 2019 inclusively. China encouraged and magnified trolling behaviour on the internet by “promoting hate speech, breaching privacy issues and malicious spreading of false information” (Richter, 2019, p. 1). The study confirmed Facebook restrictions remain in place for China, Iran and Syria, and in further oppressive control, North Korea has enforced complete internet restrictions. Australia, along with many western allies that include the United States of America and the United Kingdom, was considered to have unrestricted access to the internet. The report was critical.
of the role of social media and flagged the absence of universal responsibility and cohesive action a threat to society. “Social media allows ordinary people, civic groups, and journalists to reach a vast audience at little or no cost, but they have also provided an extremely useful and inexpensive platform for malign influence” (Shahbaz & Funk, 2019, p. 2). This global communication platform has been leveraged by trolls. It is important that this study considers these factors when analysing findings and drawing its conclusions.

Accessed by 2.603 billion people worldwide in 2020, Facebook is considered in many countries an integral part of society, playing a fundamental role in governing communication, engagement, and cultural evolution (Bonson, Royo & Ratkai, 2014). Over time, Facebook emerged as central part of citizens’ personal and cultural identity (Seru & Magogwe, 2016). The interactivity that Facebook designed to increase connectivity and transparency has triggered further segregation and social divides of censored regions. China is representative of this emerging divide, with strict government regulations meaning only 0.2% of its 1.404 billion population in 2020 have Facebook access, and only 4.1% of the population have access to other foreign media platforms including YouTube, Instagram, Pinterest and Twitter. In neighbouring Taiwan, there are no restrictions and consequently 98.9% of Taiwan’s population use Facebook. Differing political views on internet access may explain why Taiwan has a representative on the Facebook Advisory Board and China does not. The study spotlights the potential for victim impacts to intensify when experienced in areas of imposed isolation enforced by media regulations and government authorities.

On May 16, 2021, intense debate about the treatment of women on platforms such as a Facebook-like platform run by Weibo was sparked by the abuse of Xiao Meili, who posted to Weibo a video of a man who threw hot liquid at her after she asked him to stop smoking (Alto, 2021). Chinese activists who spoke out against the state became the target of a torrent of vitriol for criticising or drawing attention to human rights issues, particularly women’s rights. Feminist discourse prompted vile abuse from regular citizens online, as well as from government officials and state media (Davidson, 2021). Journalists reporting such topics were attacked,
accused of being traitors, harassed by the distribution of fake nude photographs, and threatened with harm to family members (Davidson, 2021). Researchers who analysed propaganda activities, including Vicky Xiuzhong Xu, were also targeted as China has imposed sanctions on scholars in further efforts to silence critics (Kuo & Shih, 2021). Although the scope of the study limits an in-depth exploration of this theme, anonymity is considered a key strategy to combat online abuse that limitations on internet freedom shields.

**Discriminatory leverage**

The study showed that the Facebook Advisory Board predominately represents users from westernised countries. A bias can be observed in the predominantly western resources, advice, tools and materials Facebook provides. Some of the literature that underpins discussion points included Salter, Adams, and Perez (2017), who illustrated how racism can be produced and reproduced through preference and selection. Only some resources were provided by Facebook in languages other than English, which both signals the platform's potential for linguistic bias and, by extension, cultural bias. The complexities of different cultures may not be adequately considered if support systems are viewed through a narrow and western philosophical lens. Western perspectives may not have a comprehensive understanding of cultural differences that may be targeted by trolls. Failing to recognise and address the vulnerabilities of certain cultural groups may undermine the effectiveness of Facebook's support strategies. Material provided by the Facebook Advisory Board may not be equally beneficial for all users. A more nuanced understanding of the impact of trolling on Facebook’s majority (non-white) audience could be gained by including greater diversity on the board. A more multi-cultural approach including a range of perspectives about the impacts of trolling would enrich education material and could enhance user experiences on the service.

**Affluence bias**

In addition, this study suggests developed nations may have a greater funding capacity for trolling research, and therefore be more able to effectively contribute to safety procedures. This may potentially justify the greater percentage of
western organisations represented on the Facebook Advisory Board. The literature review for this study revealed that the first academic research on trolling emerged from the United States, United Kingdom and Australia. From the English literature language available, western journalists and their publishers were among the first to speak out about trolling in their workplace, including the American-based Jezebel, which published Lindy West’s 2015 article, *What happened when I confronted my cruel troll*, that inspired this study. These examples provided some of the first case studies for investigation by academic researchers, including Australian Emma Jane (2015). This may perhaps justify the inclusion of multiple organisations from westernised countries on the board; however, it does not justify the exclusion of developing countries. The study suggests the collaborative knowledge provided by organisations from a broader cross-section of countries would improve Facebook’s responses to trolling.

**Limitation**

The study acknowledges the limitation of evaluating archival evidence published in English only. Assessing Facebook responses in languages other than English fell beyond the scope of this study. However, the study recognises developing nations not represented on the Facebook Advisory Board may have different safety services in response to trolling omitted from the analysis due to language barriers.

**Facebook Safety Centre support strategy background**

The Facebook Safety Centre was identified by the study as the second support strategy in response to trolling. To provide context for analysis, the study provided a brief background from descriptions on Facebook’s website: https://www.facebook.com/safety. Launched on April 13, 2010, the Facebook Safety Centre is an online information page dedicated to online safety and providing links to related policies, tools and resources. The page works with external experts including a Safety Advisory Board (described in the previous discussion section) to gather user feedback to assist in the development of updated response strategies to trolling on the platform. Facebook announced the page, “quadrupled the safety content available, and created a cleaner, more navigable interfaces to help users find answers to safety questions fast”
Both the appearance and useability of the page evolved over time as more interactive features were added, including video content. For a visual presentation of these changes, the study retrieved archival evidence the original 2010 interface (displayed in Figure 25 on the following page) for comparison with the most recent 2020 interface (displayed in Figure 26 on p. 151).

Figure 25

The initial interface of Facebook’s Safety Centre when launched in 2010

(Facebook, 2010)
In response to increased global usage, the Facebook Safety Centre was redesigned and translated into 50 languages on October 24, 2016, guiding users through steps to manage their interactions with other users using privacy and security tools (Lunden, 2016). Despite these translations, the study reiterates there were few non-western organisations represented on the Facebook Advisory Board that produced this content, and generic information lacked cultural sensitivity.

As an extension of the established Safety Centre, Facebook integrated the additional two resources: the Parent Portal and Youth Portal. Despite their introduction, these resources are not particularly relevant to Australian journalists seeking to access information in response to being trolled as part of
their work. They are briefly described to illustrate the adoption of a multi-levelled systems approach by Facebook.

**Facebook Parents Portal support strategy background**

Facebook Parent’s Portal was launched on December 13, 2016, and provided parents with an education tool outlining its functions, along with a step-by-step guide to sign up and connect with others on the service. The page also provided basic advice about methods to help children safely navigate social media. While the content did not exclusively focus on trolling, related material was provided including an explanation of online privacy rights and demonstrations of “block” and “report” features to limit interactions with trolls.

**Facebook Youth Portal support strategy background**

Facebook Youth Portal was launched on May 15, 2018, and aimed to provide a guide for teenagers navigating the platform. Potential risks were flagged, accompanied by educational material that suggested prevention methods in addition to coping strategies. The page promoted discussion between users of the same age, and suggested real-life experiences were more actively engaged with and understood when heard through the voices of peers.

**Facebook Safety Centre support strategy qualitative analysis**

Facebook Safety Centre addressed individual strategies to three stakeholder groups: youth, parents and educators. Evidence of these tailored strategies are identified in the original interface in 2010, and the most recent interface in 2020. Facebook’s provision of information for various types of stakeholders resembles the Systems Theory approach described by Ackoff’s in 1974. In this strategy, the Youth Portal and Parents Portal aim to protect young people from becoming victims. They seek to mobilise young people, parents and educators to be active at the support and community level of the model, and educators as curriculum developers at the governance level. This approach underpins this thesis and Facebook’s breakdown of the social problem into stakeholder groups in a cohesive and multilayered approach may indicate a theory informed attempt towards a grounded solution. This use or an integrated approach highlights the value of
theory-informed practice, less evident in Twitter’s responses. Facebook’s apparent use of theory in developing this response strategy may have contributed to public consensus that Facebook had managed the trolling problem more effectively in practical implementation than Twitter (Hildebrand & Matheson, 2012; Tiku & Newton, 2015).

**Be Bold Stop Bullying campaign support strategy background**

The *Be Bold Stop Bullying* campaign was Facebook’s third support strategy in response to trolling. The campaign could also be classified as part of the Facebook corporate approach strategy discussed later in section 4.5.5 of this chapter. A brief background is summarised to provide context for analysis in descriptions from articles about the *Be Bold Stop Bullying* App, which in 2021 no longer exists. On November 1, 2012, Facebook launched the *Be Bold Stop Bullying* campaign in Australia, following a similar petition in the United States that acquired over one million signatures in support. Facebook’s campaign encouraged users to stand up to online bullies by calling out their behaviour. The central feature of the campaign was an interactive app where adults and children pledged to rally against bullying by reporting abusive instances. Some examples of student pledges listed on the app included, “I will take a stand when I see young people humiliating or hurting each other,” and “I will not use my phone or computer to spread rumours or say hateful things, and I won’t ignore it when others are cruel and intimidating” (Marszalek, 2012, para. 4). Adults were encouraged by the app to dismiss the misconception that bullying is simply “young people being young people” (para. 4) and to actively seek intervention. Integration with multiple stakeholders was promoted, with adults encouraged to work with schools, coaches, and parents to eradicate bullying, especially if their own child was involved. The app included an interactive map displaying how many people had pledged to the campaign in each Australian state or territory and provided information on practical resources victims could access for help. Although the pledging process aligns with the support/peer/community level of Ackoff’s system model, if it was implemented on its own it could be seen as an abdication of responsibility to take action on the governance level. This is called self-regulation in policy terms, however, none of the ten participants in chapter 4.1’s interviews remembered this campaign. As
Facebook seemed to be also taking action on the governance level such as adaptations to feature strategies, it was part of a more extensive multi-levelled approach.

**Be Bold Stop Bullying campaign support strategy thematic analysis**

From the thematic analysis of qualitative data about the *Be Bold Stop Bullying* campaign in 2012, the study identified four emerging triggers for corporate responses including external pressure, accumulating negative publicity, self-governance, and misdirection, which are discussed and critiqued below.

**External pressure**

An external pressure theme was identified for discussion and critique. Facebook’s *Be Bold Stop Bullying* campaign was launched after similar campaigns were established by other news organisations and activist groups including the #StopTheTrolls campaign. Only after external pressure was exerted via critical commentary from victims, the Australian media and Australian law enforcement did Facebook react, launching its own campaign. Facebook was not a forefront leader regarding the anti-trolling campaigns described in detail in section 4.3 of this study (on p. 82). The first campaigns that spoke out about the impact of trolling and suggested response strategies to be implemented in both the public and private realm were predominately spearheaded by female journalists who expressed feminist views. However, Facebook did follow suit and launched its own campaign after other stakeholders had taken the lead.

The #StopTheTrolls campaign is a notable Australian newsroom’s response to trolling and is discussed in further detail in chapter 4.5 of this study (on p. 139). However, the campaign is again briefly mentioned as a central part of this discussion section to provide context for the external pressure mounting on Facebook before its own *Be Bold Stop Bullying* campaign in response to trolling was initiated. On December 11, 2012, the Australian publication *The Daily Telegraph* announced the #StopTheTrolls campaign. The campaign suggested the integration of a holistic approach to trolling was critical to any success by stating “what we need is strong co-operation from governments, law enforcement and the
community” (Roxon, as cited in Jones & Byrnes, 2012, para. 8). This approach mirrored Facebook’s perspective at the time, and by 2014 Facebook had introduced an “empathy team” to help its engineers and designers understand the user experience (O’Reilly, 2014, para. 3). Both campaigns appear, in their calls for action on multiple levels, to be influenced by systems theory thinking, as proposed by Ackoff (1974). The campaign gathered multiple stakeholders who put external pressure on Facebook to respond. The campaigns identified Government, law enforcement, education, non-for-profit organisations, and celebrities as stakeholder groups. In addition, the newsroom stakeholder group that launched the initiative was also a stakeholder. This study identified two stakeholder groups, the government and law enforcement, that were particularly critical of Facebook’s inaction. Their critique is unpacked below.

Government criticism emerged through comments from the then Australian Attorney-General Nicola Roxon in 2012, who called out Facebook’s inaction by stating, “we need the assistance of US-based social networks” (Jones & Byrnes, cited Roxon, 2012, para. 8). In additional criticism from the Government, Australian Communications Minister Stephen Conroy in 2012 acknowledged Facebook, unlike Twitter, had hired staff in Australia to address abuse (Hildebrand & Matheson, 2012, para. 24) but suggested further action from the platform was critical. Conroy addressed the need for the platform to disclose the identities of flagged users, stating “[the platform] may think they are above Australian law... [but] they should co-operate with law enforcement authorities to prosecute trolls” (Conroy, cited in Hildebrand & Matheson, 2012, para. 24).

Law enforcement stakeholder criticism emerged with comments from Australian police that confirmed its “limited ability to seek the identities or IP addresses of anonymous abusers from social media platforms,” (Jones & Byrnes, 2012, para. 11). Victorian Police Commissioner Ken Lay expressed disdain for Facebook one month before the Be Bold Stop Bullying campaign was launched. In early October 2012, Lay stated “the company had no social responsibility when it resisted calls to take down pages the commissioner feared would undermine the trial of the man accused of murdering Irish expatriate Jill Meagher in Melbourne” (Lay, as cited in
These political statements indicate a lack of synchronisation between social media platforms and Australian authorities, contributing to the trolls’ ability to break the law and abuse their victims without fear of prosecution. This external pressure from a widespread group of stakeholders, amplified by The Daily Telegraph’s #StopTheTrolls campaign, appears to have contributed to the shifting episteme prompted Facebook’s global Be Bold Stop Bullying campaign as the corporation’s own response to trolling.

**Accumulating negative publicity**

Accumulating negative publicity was a theme identified for discussion and critique. Disparate discussions about the presence and impact of trolling on Facebook accumulated around 2011 in the *social problem period*, and swiftly snowballed into widespread disapproval that was echoed throughout stakeholder commentary in 2012. This rising choir of negative voices reached a crescendo with the previously mentioned #StopTheTrolls campaign. In supporting evidence of the growing hostility, Australian journalist Richard Chirgwin (2012) stated, “the Be Bold Stop Bullying campaign launch gives Facebook some desperately needed favourable publicity in the Murdoch press in Australia, which had the company in the cross-hairs during its #StopTheTrolls campaign” (para. 5).

**Self-governance**

This study identified self-governance as a theme for discussion and critique. The Be Bold Stop Bullying campaign encouraged all users to download an app that invited users to pledge their alliance to cyber-kindness in the form of a petition. Users were advised to report abusive instances to Facebook, through the standard reporting functions. This study identified no new reporting features provided by Facebook to accompany the campaign. In addition, the study identified that there was no new accountability measure for users who signed up to the Facebook campaign. There was no measurable means to determine if the pledged users implemented the recommended strategies suggested by the campaign or not. The campaign did, however, provide a forum for discussion of trolling and its impact on victims, and enabled users to connect with others with a shared interest on the topic. The campaign did increase awareness on the topic of trolling (Marszalek,
2012) and provided support for trolling victims via referrals and prompted a more comprehensive understanding of trolling in the community. However, the campaign was not a methodical response to trolling, and this may have been a result of the rush to launch due to mounting negative publicity from the external pressures previously discussed. The study further suggests the campaign may have been constructed as a response to the hostility rather than a response to trolling and suggests the campaign could be more correctly labelled as a publicity tactic that future research could analyse in more depth using situational crisis communication theory (Coombs, 2007).

**Misdirection**

As discussed in the previous section, Facebook’s *Be Bold Stop Bullying* campaign did not provide any additional technological changes or functionality improvements to the service. Critics accused Facebook of directing the problem externally, rather than dealing with the true cause of the problem. Chirgwin (2012) accused the campaign of “flicking the problem to users” (para. 2) rather than strategizing ways in which to remove the problematic content from the platform. The root cause of the trolling problem may be the inability of social media companies to adequately control their own service as supported by Binns (2017), Bossio and Holton (2021), Duggan (2017) and Barnes (2018). Facebook aims to provide an online platform for interactive discussion between users; however, that online discussion, left unmonitored, is the pathway that trolls exploit to reach victims. The *Be Bold Stop Bullying* campaign does not address this key issue and directed attention away from its inability to remove abuse distributed on the platform, its inability to protect users from abuse on the platform, and its inability to provide a safe product for its users. If a physical product or service was determined to fail in these three identified areas, it may be recalled, removed, shut down until the problem was rectified or further investigated by the Australian Competition and Consumer Commission. However, because the issue emerged within the online domain, a differing set of rules and regulations apparently applied. The study identified a continued increase in Facebook usership at the time (in 2012), suggesting that users overlooked the cultural norms that would usually be associated with a physical faulty product or
service simply because the service was online. Instead, public discourse did not demand reprimand for Facebook’s inaction, and users continued to engage on the platform and passively accepted its failure to keep people safe. This made it harder for trolled journalists to leave.

**Facebook Bullying Prevention Hub support strategy background**

The Bullying Prevention Hub was identified by this study as the fourth support strategy in response to trolling by Facebook. To provide context for analysis, a brief background and summary of its role is provided below from descriptions on Facebook’s website. On November 6, 2013, Facebook launched the Bullying Prevention Hub. Developed in partnership with the Yale Centre for Emotional Intelligence, the online information page was dedicated to the prevention of harassment and provided clear and accessible resources that addressed three separate stakeholder groups: teenagers, parents and educators. Stakeholder groups were further sub-categorised by Facebook to provide tailored step-by-step plans for users of differentiating circumstances.

1. Teenage support strategies were individually provided under the headings:
   - If you’re being bullied
   - If your friend is being bullied
   - If you’ve been called a bully

2. Parent support strategies were individually provided under the headings:
   - My child is being bullied
   - My child is bullying others

3. Educator support strategies were individually provided under the headings:
   - My student is being bullied
   - My student is bullying others
   - Prevention
The layered approach to trolling is congruent with Ackoff’s (1974) System’s Theory, although it omits police, legislation, journalists and celebrities as social positions in the episteme. Teenage support strategies align with Ackoff’s self-control/individual level; parent support strategies align with the humanisation/support level; and educator support strategies align with the environmentalisation/governance level, whether it is to do with teachers providing supportive spaces for trolling victims and/or their perpetrators. These levels are illustrated in Stakeholders from the Trolling Paradigm categorised by levels in Ackoff’s (1974) Systems model in Figure 2 (on p. 5).

**Facebook Bullying Prevention Hub support strategy thematic analysis**

From the thematic analysis of qualitative data, the simplification in the *Be Bold Stop Bullying* campaign response was identified for further discussion. The step-by-step guides provided by the Bullying Prevention Hub suggested tactics to prevent online harassment and provided methods to prompt conversations with others about their experiences. While the simplicity of this information may have helped some users, inversely, it may have undermined the severity of some trolling experiences. While this seems counterintuitive it may have effected some of the most vulnerable users, such as 14-year-old Amy “Dolly” Everett, who took her own life after sustained cyberbullying. This notion was put forward by O’Driscoll (2018) in her critique suggesting, “the advice [provided in the step-by-step guide by Bullying Prevention Hub] seems largely oversimplified” (para. 45) and ineffective in practical application. For example, a scenario presented on the Bullying Prevention Hub (included in the data for this section) suggested steps for parents to take if their teenager was being threatened, but these did not include any consideration of police involvement. The central theme of parental advice was to report the offence to the victim’s school principal, which failed to address that abuse can involve children who attend different schools. Trolling incidences that originate beyond physical confinements of the schoolyard cannot be resolved through mediation between the parties in person. The globalisation of the internet has provided an avenue for transcontinental trolls, which the guides provided by the Bullying Prevention Hub do not address. In further evidence of the ineffectiveness of the suggestions offered, teenagers who have been accused of
bullying are advised to apologise to their victim. Although apologies are not necessarily bad ideas, the advice does not consider whether the severity of abuse may constitute a crime. For these cases, perpetrators should not be encouraged to seek out and apologise to the victim, but instead turn themselves over to law enforcement or seek professional treatment, such as psychotherapy.

Facebook’s public figures policy was introduced in 2019 to encourage user engagement with creators, authors, athletes, sports teams, musicians, bands, journalists and celebrities (Smith & Goldman, 2019). However, the policy allowed more critical commentary of public figures than of private individuals (Paul, 2021), which in 2019 included journalists. In 2021, Facebook modified its public figures policy to specifically exclude journalists, with Facebook Global Head of Safety, Antigone Davis, labelling journalists as “involuntary” public figures (Davis, 2021, para. 3). Facebook changed its approach on the harassment of journalists and now, as of 2021, considered journalists to be in the public eye because of their work rather than their public personas (Paul, 2021). Diaz (2021, October) suggested the response was not triggered by activism, but in response to criticism in the wake of whistleblower Frances Haugen’s interview and Congressional testimony stating, “as long as Facebook is operating in the shadows, hiding its research from public scrutiny, it is unaccountable” (Shepardson & Bartz, 2021). Although Facebook’s public figures policy was modified to exclude journalists and therefore protect them from trolls, anecdotal evidence suggests it has not eliminated the trolling of journalists on the platform (Price, 2022).
4.5.2 Facebook feature strategies

Feature strategies were identified by the study as the second type of Facebook response to trolling. The study identified sixteen feature strategies that Facebook implemented on its platform in response to trolling, derived from the quantitative and qualitative data results of this section. A complete list of every feature change was not feasible, as documented archives available from Facebook.com were incomplete, and the meticulous investigation required to independently produce such an extensive inventory was not within the scope of the study. However, it provides a comprehensive list of features, from its launch in 2009 to the date of analysis in May 2020. Information from Product Policy meeting notes that directly addressed “keeping abuse off the service” (Facebook, 2018b, para. 1) at fortnightly conferences were a significant contribution. The sixteen features that can be actioned by Facebook users in response to trolling are listed below, and are individually described later in the discussion section:

1. Hide comment
2. Snooze
3. Unfollow
4. Take a break
5. Mute
6. Remove user from Page
7. Remove user from Group
8. Delete comment
9. Unfriend
10. Ban
11. Disable posts
12. Profanity filter on Pages
13. Blocking words on Pages
14. Disable comments on private accounts
15. Block
16. Report
These sixteen features provided qualitative data for further thematic analysis. While recognising while no one solution is likely to work alone, to assess the potential effectiveness of each feature in protecting trolling victims from abuse on the platform, nine favourable attributes were identified. Attributes were defined as favourable if they were potentially useful to journalists seeking to manage trolling on Facebook. The study first compared which of the sixteen application features had the highest number of favourable attributes; however, the results were inconclusive as application features had similar number of favourable attributes associated with them, as displayed in Figure 27 (on p. 165). Further thematic analysis methods were required to expose and interpret meaningful data.

Through a deductive process, each individual attribute was assessed to determine its relative significance from the perspective of hypothetical trolled victims formulated as an amalgam of the victims interviewed in the research presented in section 4.2 of this study (on p. 64). Nine attributes on a scale from least favourable to most favourable were identified. The study noted that abuse was often perpetrated on Facebook two types of community pages including Groups (used for people with common interests) and Pages (used for promotion of organisations, businesses and public figures) that publications such as the ABC fall under. The study recognised abuse published in these public spaces was more problematic for a victim as the authority to remove content lay with the administrators of the Group or Page. While journalists have the authority to control trolling content on their own Page, and to some extent their newsroom’s Pages, many media organisations have multiple accounts on Facebook with different moderators. In addition, not all newsroom accounts are actively moderated, even after the legal change resulting from the Dylan Voller case which is discussed in more details in the legal chapter 4.7 of this study. Journalists can also not control trolling on trolls personal Pages or the Pages of rival news organisations (Price, 2022). The study, therefore, determined the authority to implement the feature was the most influential favourable attribute and based on this notion, ranked the nine attributes in descending order of importance:
1. The victim has the authority to implement the feature.
2. All trolling content is deleted indefinitely.
3. The feature is actioned immediately.
4. An external moderator is not required.
5. An autonomous computing algorithm is utilised.
6. The troll cannot reach the victims again.
7. The feature is proactive deterrent.
8. The troll would be unaware that the victim implemented the feature.
9. The interactive function of Facebook is not removed.

For a visual representation, the findings are presented in Figure 27 (on p. 165) with the favourable attributes tiered from greatest impact in displayed in red, to least impact displayed in pale yellow.

In addition, the study noted that abusive comments can be published on the troll’s own personal profile, and therefore the authority to remove the content lies with the abuser. In this instance, the study identified the only favourable Facebook feature is ‘report’. Although the time required from actioning the report feature, via external moderation, to eventual removal content is a critical adverse attribute. However, the report feature does not exclude any user, and was able to be implemented by all stakeholders, including private users (both victims and observers), Group administrators and Page administrators. While the report feature was therefore determined to be the most favourable Facebook feature in response to trolling, two academic reports criticised Facebook’s inability to respond appropriately to user reports of abuse. In 2019, a report by Amnesty International UK found that more than half of reports lodged by women about harassment on Facebook were met with no action from the social media company (Noor, 2019). In addition, a 2021 report by researchers at the University of Sydney who looked at Facebook’s response to hate-speech in Asia, recommended that:

- Facebook extend its consultation with protected groups on their experience and management of hate speech
• develop and publicise its trusted partners channel, so that individuals and organisations have a direct hate speech reporting partner for crisis reporting issues.
• hold an annual regional hate speech roundtable for stakeholder groups, and
• recognise the role of page administrators as critical gatekeepers of hate speech content, supporting their improved regulatory literacy via training and education. (Sinpeng et al, 2021)
A table to show the tiered thematic analysis based on the 56 documents that mentioned Facebook’s responses to trolling derived from chapter 4.5 data

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Hide comment</th>
<th>Snooze</th>
<th>Unfollow</th>
<th>Take a break</th>
<th>Mute</th>
<th>Remove user from Page</th>
<th>Remove user from Group</th>
<th>Delete comment</th>
<th>Unfriend</th>
<th>Ban</th>
<th>Disable posts</th>
<th>Profanity filter on Pages</th>
<th>Blocking words on Pages</th>
<th>Disable comments on private accounts</th>
<th>Block</th>
<th>Report</th>
<th>Total number of features using each attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>Victim has authority to implement feature</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>10</td>
</tr>
<tr>
<td>All trolling content is deleted indefinitely</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>❌</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>5</td>
</tr>
<tr>
<td>Feature is actioned immediately</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>❌</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>12</td>
</tr>
<tr>
<td>External moderator is not required</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>9</td>
</tr>
<tr>
<td>Autonomous computing algorithm utilised</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>2</td>
</tr>
<tr>
<td>Troll cannot reach victim again</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>3</td>
</tr>
<tr>
<td>Proactive feature</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>❌</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>❌</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>4</td>
</tr>
<tr>
<td>Troll is unaware victim used feature</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>❌</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>8</td>
</tr>
<tr>
<td>Interactive function continues</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>❌</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>14</td>
</tr>
</tbody>
</table>
| Total number of positive feature attributes for each feature               | 5            | 5      | 5        | 5            | 1    | 1                     | 1                     | 5             | 5        | 1   | 3             | 5                          | 5                        | 5                                | 6       | 5      | --------------------------------------------
A table to show the tiered thematic analysis based on the 56 documents that mentioned Facebook’s responses to trolling

The findings revealed the continuation of the interactive function was the most frequent attribute that 87.5%, 14 of the 16, of the features employed. This feature was classified as the least important attribute, and highlights that the frequency of application is not directly proportional to its efficacy in protecting a user from abuse. The findings revealed immediate action was the second most frequent attribute that 75%, 12 of the 16, of features encompassed. However, when observing the attribute in isolation, the data reveals the pre-determined most influential report feature does not incorporate this attribute. The study, therefore, considered both the frequency of favourable attributes and the tiered ranking of each favourable attribute as a framework to deduce the feature favourability order, described in the order that Facebook introduced them as follows:

**Least favourable Facebook features examined**

1. **Hide comment (reactive)**
   A victim can choose to hide an abusive comment. This means the user cannot see the particular trolling content after initially seeing and clicking hide; however, the content is still visible to the troll, and any users who are friends with the troll on the platform. The troll would be unaware the comment has been hidden. The hide comment feature can be actioned by private users, Group administrators and Page administrators on Facebook. If the comment is on the victim’s private account, the victim has the authority to apply the hide comment. For trolling content perpetrated on a Group or Page, the respective administrators have the authority to moderate the abusive content, not the victim.

2. **Snooze (reactive)**
   A victim can choose to snooze an abuser. This means any content the abuser posts would not appear on the victim’s newsfeed for thirty days. After this remission period, the profile returns to standard notification settings without the abuser being notified of the process. The snooze feature can be actioned by the victim.
3. **Unfollow (reactive)**

A victim can choose to unfollow an abuser. This means any content the abuser posts would not appear on the victim’s newsfeed. The unfollow feature can be actioned by the victim.

4. **Take a break (reactive)**

A victim can choose to take a break from an abuser. Firstly, the equivalent functions of the unfollow feature are applied whereby the victim can no longer see the abuser’s posts on their feed. Secondly, the abuser can no longer view the victim’s private posts. Thirdly, the victim is provided the additional privilege of deleting past posts with the abuser. This feature was aimed to provide a safety net for the ease of linked content to be removed for users in a relationship; however, it can be utilised for anti-trolling purposes. The take-a-break feature can be actioned by the victim.

5. **Mute (reactive)**

Group administrators can temporarily stop an abuser from posting or commenting in their Group. The victim does not have the authority to apply the mute feature. The mute feature can be actioned by Group administrators only.

6. **Remove user from Page (reactive)**

The ‘remove user’ function has different capabilities for Pages and Groups and they have therefore been separated for isolated discussion and ranked accordingly. Page administrators can remove an identified abuser from liking their Page. However, Pages are public spaces and removed abusers can choose to like the Page again. The victim does not have the authority to apply the feature, which can only be actioned by Page administrators.

**Favourable Facebook features examined**

7. **Remove user from Group (reactive)**

Group administrators can remove an abuser from their managed Group indefinitely. Removed members must request to join the Group again, which can
be denied. The victim does not have the authority to apply the feature, which can only be actioned by Group administrators.

8. Delete comment (reactive)

A victim can choose to delete an abusive comment. This removes the particular comment from the platform. The delete comment feature can be actioned by private users, Group administrators and Page administrators. The victim has the authority to apply the feature to moderate the abusive content if perpetrated on their personal profile. The victim does not have the authority to apply the delete comment feature to moderate the abusive content if perpetrated on a Group or Page, their respective administrators do.

9. Unfriend (reactive)

A victim can choose to unfriend an abuser. This means the abuser can no longer comment on any of the victim's private content. The unfriend feature can be actioned by the victim.

10. Ban (reactive)

Page administrators can remove an abuser from a Page. The banned abuser is still able to share Page content, but can no longer like, comment, message or publish on the Page. The victim does not have the authority to apply the ban feature. The ban feature can be actioned by Page administrators only.

11. Disable posts (proactive)

Page administrators can choose disable posts on their respective Page. The victim does not have the authority to apply the feature unless the victim is the Page administrator. The study recognised the main function of a Page is to promote a business or public figure. When a Page applies the disable-posts feature, its interactive function is completely removed. However, this interactive function is important for community reach and engagement, which increases the popularity of the product, service, or person that the Page promotes. Acting as a double-edge sword, this interactive function is critical for the news media and journalists. The study determined the feature detracts from a Page's main objective to promote
and took this reduced functionality into consideration when classifying the overall favourability of the feature.

Most favourable Facebook features examined

12. Profanity filter on Pages (proactive)

Page administrators can choose to block a varying degree of profanity from appearing on their Pages. The profanity filter is set at ‘general’ by default and can be increased to medium or strong. Once selected, Facebook’s autonomous algorithms instantly remove any post or comment that contains words and phrases determined by Facebook that the community has previously marked as offensive. Facebook continues the external moderation of the Page and independently updates this process without the need for further input from Page administrators. Profanities, racist expressions and hate speech are often blocked with this feature. Victims do not have the authority to action the feature. In addition, other users do not have an input on the level the page sets its profanity filter. Some Page administrators may be unaware of the filter, or not know how to apply the feature that is not turned on by default. In these instances, its function as a feature to combat trolling is redundant. The reliance on Page administrators to implement the feature is a significant limiting factor. These collective obstructions were taken into consideration when classifying the overall favourability of the feature.

13. Blocking words on Pages (proactive)

In addition to the profanity feature described above, Page administrators can choose to block particular words or phrases from any post or comment published on the Page. This enables certain groups to combat particular content that may be inflammatory to the cluster of users frequenting a particular page. The feature allows Page administrators to individualise a strategy against content that has the potential offend their audience. Functionality restrictions require the input of both singular and plural form of selected words for exclusion. In addition, the exact spelling of each selected word can be excluded, slight variations in lettering order by abusers may allow for the content to pass through. For example, abusers
who replace the letter E in a word with the number 3 can bypass the algorithm and the abusive content is still decipherable to users. The same limiting factors that apply to the profanity feature, equally apply to the blocking words feature. The study acknowledges that feature is only useful if Page administrators are aware of the filter, know how to apply it successfully and, importantly, choose to apply it. Otherwise, its function as a feature to combat trolling is redundant. These collective obstructions were taken into consideration by the study when classifying the overall favourability of the feature.

14. Disable comments on private accounts (proactive)
A user can choose the disable comments feature that usually appears below a post. This means no user, or potential abuser, can comment on the post. The victim or potential victim has the authority to select the remove-commenting feature for any post on their private account. Additionally, victims have the authority to apply the remove-commenting feature on posts they publish on any Group or Page. The abuser still has options to target the victim including commenting on other posts without the commenting function removed, commenting on profile photos, and privately messaging the victim. The study therefore classified this function as a most favourable feature due to its proactive functionality, but less favourable than the block feature (detailed next), as it does not block the abuser from contacting the victim entirely. Like the disable-posts feature, all interactive functions are eliminated. The disable comments feature does not allow users to converse about the particular topic, excluding all potential commentary, including the constructively critical or supportive. The study noted that although the victim had the authority to implement the feature, the reduced interactive function limited communication with all users, not just potential abusers. If all posts implemented this feature, Facebook would operate like a website rather than a social media platform.

15. Block (reactive)
A victim can choose to block an abuser. This means the abuser can no longer view any past or future content posted by the victim. The victim has the authority to action the block feature. In addition, the feature can be applied by those who
witness the abuse but are not directly targeted, including other private users, Group administrators and Page administrators.

16. Report (reactive)
A victim or observer can choose to report an abuser. The victim has the authority to initiate the report feature. However, the subsequent outcome of the feature is subject to appraisal by Facebook’s moderation team. In addition, the feature can be applied by those who witness the abuse but are not directly targeted, including other private users, Group administrators and Page administrators. Once an abuser is flagged and deemed to have breached Facebook’s Community Standards, the content is removed, and the abuser may receive a temporary ban from the platform. For multiple offences, the user may receive a permanent ban from Facebook.

Facebook feature strategy thematic analysis
From the application of thematic analysis of qualitative and quantitative data results, three themes emerged: authority, greater user control, and detailed reporting process. These are discussed and critiqued below.

Authority
Facebook provided users with the ability to independently implement operational tools such as the ‘delete comment’ and ‘block’ feature to immediately remove abusive content from personal accounts without the requirement for an intermediate moderator to access the content first. Although Facebook claims the process “usually within 24 hours” (Facebook, 2021, para. 5), critics claim “a typical time frame is anywhere from 48 hours to 45 days” (Webb, 2021, para. 5). This functional ability did not prevent trolling content being received by victims on Facebook; however, it did give users the ability to affect its instant removal, without the need to seek authority from a Facebook moderator or Page admin.

The disable-comments feature provides users with greater content control. Facebook provides the victim with the authority to remove comments, eliminating the most common avenue trolls use to target victims. Support of the feature
change is echoed in the structural changes on the websites of the many Australian newsrooms that have also removed comment streams previously featured below all published articles with the ABC in 2017 and Mumbrella in 2020. Comment stream removal is described and analysed in the 4.6 newsroom responses chapter (on p. 219). While reducing defamation risk, the study acknowledged that the removal of the interactive comment stream function negatively impacts content circulation and limits potential connections between posts and readers, who often share content by tagging other users in comments sections. Balancing the benefits of interactive functions with the detriments of trolling has proved problematic for moderators, including Group and Page administrators. Third-party mediators employed by newsrooms, whose objectives are to increase or sustain user engagement, may be less inclined to implement the disable-post and disable-comments features in comparison to a targeted victim whose main objective is safety from trolling. As Facebook users often have the authority to action its features, the power balance is regularly tilted in the victim’s favour and provides a deterrent for trolls, many of whom moved to platforms such Twitter that has been regulated less. This suggestion is further unpacked in the discussion and critique of the next theme.

**Greater user control**

From Facebook’s initial launch in 2004, users have had more control over published content in comparison to other social media platforms such as Twitter and Instagram. Although Facebook implemented the delete feature from its launch date in 2004, 17 years later, in 2021, Twitter users still do not have the authority to remove abuse and must report such content to an external moderator for review. As discussed in chapter 4.4 (on p. 101) Twitter’s continued failure to provide users with control has created a cesspit of abuse on the platform and lead to many journalists advocating to cease use of the service (West, 2015; Hobson, 2019; Lieberman, 2020; Kannan, 2021). This study identified that the functionality difference between the two platforms safeguarded potential trolling victims on Facebook, but left Twitter users exposed. This precise feature may be the determining factor that contributed to trolling flourishing on Twitter, and its decreasing impact on Facebook by comparison.
Detailed reporting process

Through thematic analysis of Facebook responses, Facebook’s detailed reporting process emerged as a theme for discussion and critique. From its launch in 2004, Facebook’s reporting process required victims to detail their experience through multiple-choice questions, providing descriptive data for Facebook moderators to respond to each individual complaint more efficiently. In addition, the detailed reporting process yields more data for collective analysis which can be used for generalised response strategies to trolling by the platform (Facebook, 2018b). With more data comes a greater ability to identify triggering factors in the behaviour of trolls, that can potentially enhance detection and removal of them from the service.
4.5.3 Facebook functional strategies results and discussion

Functional strategies emerged as the third Facebook response to trolling. The study identified two functional strategies (and two sub-categories) that could assist Australian journalists trolled on the platform. The qualitative data results of this section are itemised below and are individually described later in the discussion.

1. Personalised sign-up process

2. Machine learning and artificial intelligence
   
   i. Image detection algorithm
   
   ii. Language detection algorithm

Emerging themes within each of the two identified functional strategies (and two subcategories) are individually analysed, discussed, and critiqued in the following sections.

Personalised sign-up process functional strategy background

From the 56 articles retrieved as archival evidence in this chapter, the personalised sign-up process was identified as the first Facebook functional strategy in response to trolling. Through exploration of the personalised sign-up process in 2020 on Facebook’s website https://www.facebook.com, the study identified the minimum amount of information required to create a personal Facebook account was a first name, last name, valid email address, password, gender and date-of-birth. Although a profile picture was not a requirement of the sign-up process, the platform strongly recommended one to be uploaded. A cover photo was also not a requirement of the sign-up process, but again the platform strongly recommended one to be uploaded. However, there was no identification requirement for the profile picture or cover photo allowing uploaded pictures to be random images. On June 15, 2012, Facebook added a phone number verification requirement for all users, which included new sign-ups and...
Previously existing accounts (Constine, 2012). The personalised sign-up process was created to minimise fake accounts being created on Facebook. The study identified that many victims were abused by fake accounts used by trolls to avoid any personal repercussions of their abusive behaviour including Phillips, 2015; West, 2015; Ford, 2016; Jones, 2017; Barnes, 2018; Gorman, 2019. However, while the personalised sign-up process proved a hindrance to the formation of trolling accounts on Facebook, it was not a complete prevention strategy.

**Personalised sign-up process functional strategy thematic analysis**

From the thematic analysis of documents published by Facebook between 2009 and 2019 about its policy updates to trolling, the personalised sign-up process functional strategy, four themes emerged: fake accounts still permeate the platform, some trolls use personal accounts; authentication tools can be exploited; and there may be potential ulterior motives for phone number verification. These themes are discussed and critiqued below.

**Fake accounts still permeate the platform**

While Facebook’s personalised sign-up process was more detailed than other social media platforms, including Instagram (of the same ownership) and Twitter when its phone number verification was introduced in 2012, the fact that fake accounts still permeate the platform with Facebook deleting 3.2 billion fake accounts in November 2021 (Rana & Paul, 2021) suggests the strategy is not a comprehensive solution to deter trolls. More deliberate attempts to verify users began to address concerns raised by academics that anonymity on social media facilitated a culture of transgression and impunity in which abuse thrived (Salter, 2016). Although verification measures reduced the number of fake accounts being created, the process was not impermeable and pseudonymous accounts used to abuse journalists still infiltrated the platform. The issue was spotlighted in campaign efforts on March 12, 2021, that called for legislation that made a verified form of identification a mandatory element when creating a social media account (Hutchinson, 2021). Continued public discourse about the verification process on Facebook and other social media platforms highlights the issue as a significant concern in contemporary journalism.
Some trolls use personal accounts

As established in the literature review and supported by findings from interviews with Australian trolled journalists in chapter 4.2 (on p. 64), trolls do not exclusively use fake accounts to troll victims, and frequently engage in the abuse of victims under their own names and identities. Some trolls actively seek to publicise their abusive content on the platform and use their public identity to increase exposure. Salter (2016) further supports this finding and suggests, “the same publicity that is turned against the victim... is maximised to the benefit of the perpetrator for whom the abuse acts as a kind of public and interactive performance” (p. 51). This identity verification as a combative strategy is therefore ineffective towards this group of abusers who reveal their identity.

Authentication tools exploited

Some of the authentication tools implemented to minimise the creation of fake accounts can be used to target victims. Trolls are able to exploit the requirements for profile pictures to identify minority groups based on appearance and cultural beliefs (Mantilla, 2013). In the study’s previous 4.2 impacts response findings (on p. 64) thematic analysis of these in-depth interviews revealed five out of the ten interviewees received appearance-based trolling abuse on Facebook. This appearance-based trolling was specifically facilitated by Facebook's profile picture. Without a profile picture, abusers would have less grounds to attack another user's physical characteristics, however, a journalists' sex and racial background can potentially be derived from by-lines or by their voice if on radio or television. Although the study acknowledges the photograph is not a requirement of Facebook's sign-up processes, it does recognise that Facebook strongly recommends and encourages all users to participate and upload a personal photo. Furthering this, if an account is reported, even if incorrectly so, Facebook requires the user to upload personal identification details to the service to match their corresponding profile picture (Tiku, 2017). If the user does not have a profile picture, the moderation team has no comparison, and the account may not be cleared. Profile pictures are optional for any account that is not flagged; however, when brought under question to resolve a dispute, as trolling commonly entails, one is required for the trolls. However, one participant from the interviews
conducted in chapter 4.2 of this study (on p. 64) revealed they were trolled because of their appearance displayed in their profile picture. Trolls can also exploit usernames that are commonly associated with certain ethnic regions to discriminate and target specific groups of users for racist-based and religious-based abuse. However, the benefits of reducing the number of fake accounts on the platforms by implementing the authentication requirements seem to outweigh its potential exploitation of the function by trolls.

**Potential ulterior motives for phone number verification**

The motives behind the introduction of the phone number verification requirement were questioned, with some critics suggesting the strategy aimed to “cash in on security fears” (Constine, 2012, para. 1) rather than to combat trolling content on the service. In the wake of the Cambridge Analytica scandal, which revealed that at least 87 million Facebook users had their public data used without explicit consent (Gilbert, 2018) to influence elections including the 2016 US Presidential election, critical debate emerged, placing the financial motives of the platform in public discourse. However, in 2018, Facebook founder Mark Zuckerberg declared “there’s a very common misperception about Facebook – that we sell data to advertisers… we do not sell data to advertisers. We don’t sell data to anyone” (Zuckerberg, as cited in Gilbert, 2018, para. 3). This study acknowledges critics’ suggestions that Facebook could potentially on-sell user data and utilise the information as a potential revenue stream; however, Facebook refutes this and maintains that the primary objective of implementing phone number verification was for user security and subsequent reduction in fake accounts, recognised facilitators of abusive content.

**Machine learning and artificial intelligence functional strategy background**

This study first defined these two key terms from descriptions in the Oxford English Dictionary (2021). Machine learning was defined as “the use and development of computer systems that are able to learn and adapt without following explicit instructions, by using algorithms and statistical models to analyse and draw inferences from patterns in data”. Artificial intelligence was defined as, “the theory and development of computer systems able to perform
tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages”. Machine learning is, therefore, a part of artificial intelligence; however, artificial intelligence involves deeper processing that can attempt more complex problem solving. Although a distinction between the two terms is important, the study recognised they work in concert, as machine-learning models provide the building blocks that inform artificial intelligence processes.

Facebook first used machine learning when it introduced the News Feed function on September 5, 2006 (Manjoo, 2013). Although identifying faces is a relatively simple task for humans, it had been a difficult task for computers to perform. By integrating facial recognition technology on its platform, Facebook could suggest friends that users could tag when uploading images (Hutchinson, 2015). By March 17, 2014, Facebook’s DeepFace project announced its image facial recognition models could identify a human face from 800 million photographs in less than five seconds with 97.25 per cent accuracy (Oremus, 2014). But it was not until June 15, 2015, with the launch of its photo-sharing service Moment, that Facebook began openly talking about how its deep research into machine learning was influencing new products (Higginbotham, 2016). Privacy concerns began to be raised, as the new technology forced users to confront the unsettling fact that Facebook could identify them out of more than a billion users in an instant. These data-driven capabilities allowed Facebook to make its platform easier to use. But they also enabled the company to keep people using its platform, which in turn allowed it to sell more, and more effective, advertisements to users. To do this, Facebook ran tens of trillions queries per day that made about six million predictions per second (Higginbotham, 2016). These trillions of data points train News Feed algorithms within hours, with learning models updated every 15 minutes to two hours in order to react quickly to current events (Higginbotham, 2016). That computers are able to analyse such vast amounts of information and make independent judgements, is a disconcerting reminder that every aspect of one’s digital life is being autonomised and dissected in ways that show advertisers, researchers, and even governments a picture of our private thoughts and actions.
Machine learning and artificial intelligence functional strategy data results

Technological advancements fuel the ability of machine learning and artificial intelligence algorithms to better detect trolling images. This progression has been gradual but, in recent years, significant progress has enabled companies such as Facebook to embrace and attempt to implement the technology. It is important to outline a brief history of this progression to provide the qualitative data for analysis and comparison. The study analysed 56 articles retrieved as archival evidence in this chapter as a starting point and drew supplementary data from articles identified by using Google as a search tool for the terms “machine learning” + “algorithm” + “artificial intelligence”. Key milestones of the evolution of machine learning and artificial intelligence are listed below.

Algorithms surpassed humans at classifying the content of images on February 4, 2015.

- Microsoft announced that its algorithms had surpassed humans at classifying the content of images (He et al, 2015).
- Critics suggested the test was calculated in a way that favoured machines (Thomsen, 2015).
- An example of this bias included a question to sort 1,000 photos into 120 dog breeds, a process better suited to machine recognition.
- However, the function fell short when required to understand and interpret the meaning of the image, beyond a basic sorting process. Social issues such as trolling are by nature complex, and it is this very complexity where artificial intelligence failed (Simonite, 2018).

Artificial intelligence speech recognition was as claimed to be as good as humans on October 18, 2016.

- Microsoft reported the standardised word-error rate of professional transcriptionists was 5.9 percent (Xiong et al, 2016).
- Microsoft reported its speech recognition system had a word error rate of 5.9 percent (Xiong et al, 2016).
• Microsoft claimed its artificial intelligence program could now perform making the same or fewer errors than professional transcriptionists (Linn, 2016).
• The announcement was labelled an “historic achievement” (Simonite, 2018, para. 6).

Artificial intelligence speech recognition was as confirmed to be as good as humans on August 20, 2017.
• Shortly after Microsoft’s previous announcement, IBM reported the accurate word-error rate of professional transcriptionists was actually 5.1 percent (Saon, 2017).
• Microsoft acknowledged the criticism and reported its speech recognition system had reached a word error rate of 5.1 percent (Xiong et al, 2017).
• Microsoft’s artificial intelligence program was confirmed to now perform making the same or fewer errors than professional transcriptionists (Huang, 2017).

Artificial intelligence matched or outperformed humans at reading and answering questions about text on January 15, 2018.
• Microsoft announced it had produced software that matched or outperformed humans on a standardised reading and comprehension test conducted by Stanford University (Linn, 2018).
• The artificial intelligence software was required to answer 10,000 simple questions about excerpts from Wikipedia articles. Researchers built their software by analysing 90,000 sample questions, with the answers attached (Simonite, 2018).
• Questions including, “where do water droplets collide with ice crystals to form precipitation?” had to be answered by highlighting words in the original text, in this case, “within a cloud” (Simonite, 2018).
• In early January, Microsoft and Alibaba submitted models that respectively got 82.65 and 82.44 percent of the highlighted segments exactly right. This was the first artificially intelligent software that had outperformed the
82.304 percent score Stanford researchers had termed “human performance” (Linn, 2018).

- However, Professor Percy Liang, one of the test’s creators, labelled the test as a “fairly narrow” evaluation of reading comprehension, and critics suggested that the test was not an accurate measure of English language proficiency.

- Engineers described the key to its success was the application of a “deep learning” approach that trained algorithms with example data (Simonite, 2018, para. 9).

- The artificial intelligence software combined Long Short-Term Memory and WaveNet language models with three strong acoustic models (Saon, 2017).

- The announcement was labelled a “major milestone” and media coverage at the time suggested millions of jobs were at risk, however, content moderators welcomed its potential to perform their current roles (Linn, 2018).

Emerging from the exploration of machine learning and artificial intelligence using the grounded theory approach, two subcategories of image detection algorithms and language detection algorithms were considered relevant to Australian journalists who were trolled on Facebook as part of their work and were selected for further inquiry.

**Image detection algorithm background**

The study explored the image detection algorithm next, using the 56 articles retrieved as archival evidence in this chapter as a starting point, and supplementary data from articles identified by using Google as a search tool for the terms “image detection” + “algorithm” + “Facebook”. Although image detection algorithms are considered a successful function strategy by Facebook and its critics, the study notes the algorithm does not remove all offensive imagery. In particular, the distribution of non-consensual intimate imagery is still prevalent on the platform, and a noted concern of the management team. Non-
consensual intimate imagery is addressed later as part of the Facebook research strategies (on p. 196).

The following statistics were retrieved from Facebook’s 2018 community standards enforcement report to analyse and interpret the effectiveness of the image detection algorithm. The study identified and calculated:

- In one year, 107.6 million images were flagged to contain adult nudity or sexual activity, and were subsequently removed from the platform, between October 1, 2017, and September 30, 2018.
- For the same time period, the image detection algorithm had an average success rate 95.75 percent for identifying adult nudity or sexual activity violations before being reported by users.
- This meant the algorithm failed to detect 4.78 million images during that year.

For comparison, the study further identified and calculated:

- For the same time period, 8.7 million images were flagged to contain child nudity or sexual exploitation and were subsequently removed from the platform.
- The image detection algorithm had an average success rate of 99.2% for identifying child nudity or sexual exploitation violations before being reported by users.
- This meant the algorithm failed to detect only 69,600 images during that year.

Due to the vast quantity of content published on Facebook, seven low rates of failure result in disturbing volumes of undetected content.

**Image detection algorithm thematic analysis**

Image recognition is important for journalists as trolls often send pornography as part of their attacks. Through thematic analysis of qualitative and quantitative data results of image detection algorithms that help to protect Australian journalists from being trolled, three themes emerged: time lag of removal;
successful implementation; and determining consent. These themes are explored and unpacked in the following sections.

**Time lag of removal**

The reliance on the reporting process was inefficient in addressing harm in a timely fashion as described by Martin and Murrell (2020). Jakubowicz et al (2017) stated, “the reporting process inherently involves a time lag within which the abusive material may be further disseminated” (p. 57) and they are particularly critical of the strategy’s effectiveness in response to cyber racism. The study notes this key failing of the reporting process and calls for interactive providers to improve responsiveness. Facebook (2019) also reported that the corporation had acknowledged the time lag between a victim reporting a trolling incident to Facebook, and a response from Facebook being received an issue flagged as a priority in 2019. Facebook initially reported that in this time period, “victims were filled with anxiety and helplessness” (para. 7). Victims required fast but personalised responses that were not provided by Facebook, as the damage they experienced increased the longer the trolling content remained online.

**Successful implementation**

Facebook claims that the image detection artificial intelligence program was a successful tool in response to trolling. This endorsement suggests that it will continue to be an important combative strategy for Facebook in the future. Microsoft was the leader that paved the way and provided the tools social media companies could choose to recognise and implement (He et al, 2015). Facebook was first social media platform to actively embrace image detection technology and, in 2014, led by Tanton Gibbs, transitioned to automated content moderation. Academics such as Simonite (2018) first criticised Facebook for assigning the job to Gibbs, an ad executive who had no expertise in online discourse at the time. However, by embracing deep learning the technology became much more versatile (McDermott, 2018). In comparison, Twitter was far slower to undertake the required back-of-house changes, initially resisting the proposed benefits of artificial intelligence advancements until following Facebook’s lead in 2019 (Leetaru, 2019).
Facebook’s success was attributed to deep learning algorithms that have proven to be effective when sorting images into categories with two alternative outcomes, such as pornographic and not pornographic. The algorithm attributes a ‘yes’ or ‘no’ value to each image, from which further functions can be applied in a systematic filtration system. However, its application is not perfect, and the algorithm been criticised for misclassifying images of nursing mothers as pornography (Bell, 2018). Identified flaws are frequently modified and refined in the development process. Early stages of success have been used as talking points for Facebook executives touting the potential of artificial intelligence to eliminate trolling on the service (McDermott, 2018). Image detection algorithms provide the first evidence that an algorithmic immune system could help shelter Facebook users from harmful content, and subsequently shield the company from the potentially harmful consequences of hosting it. Although artificial intelligence has proved effective at minimizing nudity and pornography on the platform, recognising and decoding language has proven to be a far more complex problem. The vast majority of abusive trolling content is verbal and hate speech and bullying have proven a far more difficult task to combat.

The study identified the interpretation of imagery and text is still a problematic given the limitations of current technology (Mitchell, 2020). As forecast in 1986 by mathematician and philosopher Gian-Carlo Rota who questioned, “I wonder whether or when artificial intelligence will ever crash the barrier of meaning” (Rota, 1986, p. 1), algorithms are not yet proficient in the recognition of meaning, and therefore cannot yet be universally applied against trolling. Professor Melanie Mitchell from Portland State University outlined this limitation in her research and described modern computing software as stuck behind “the barrier of meaning” (Mitchell, 2020, p. 1). Computing engineers and research teams globally are trying to figure out this critical problem; however, industry leaders, such as Microsoft, suggest that with perseverance, artificial intelligence will evolve to overcome this hurdle in the near future (Linn, 2018).
Determining consent

Explicit images are commonly published on the platform with consent of all parties. These include swimwear and lingerie promotional images, and implied-nude artistic photos of professional models and private users. Commonly permitted by Australian advertising standards (as outlined in section 2.4 of the Australian Association of National Advertisers Code of Ethics), and largely accepted by western society as inoffensive, these images are widely distributed on the service, with younger generations emerging as their most frequent sharers. Journalists have been trolled with such images including Siobhan Moyes in 2016 (Thompson, 2016) and Australian journalist Lily Mayers in 2020 (Gillespie, 2020). Such images would not be removed by the image detection algorithm as they do not breach Facebook’s Community Standards in regard to nudity or sexual content. However, western social mores do not apply to all users and many women, particularly of more conservative religious beliefs, may consider personal photos being publicly posted as extremely offensive. User consent is the influential factor when defining an image as offensive. The study identified the intention to offend was the key factor when determining if an image was published as an act of trolling or not. This study identified that the image detection algorithm is not able to determine the consent of the subject of an image, or its subsequent intent to offend, highlighting a flaw in the automation process that trolling content can potentially bypass. The subjective nature of image detection, in particular in relation to non-consensual intimate imagery, is highlighted as a significant issue and is discussed in further detail in the Facebook research strategies (on p. 196).

The study suggests removal of the consent variable may contribute to discrepancies between the rates of removal between images of nude adults and children. The image detection algorithm had an average success rate 99.2 percent for identifying child nudity or sexual exploitation violations before being reported by users, in comparison to an average success rate 95.75 percent for identifying adult nudity or sexual activity violations before being reported by users. Whether or not a child consents for the image to be posted or not is irrelevant, as the act is illegal, and the image would be immediately removed. The consent variable would only apply to adult imagery and, due to its subjective nature, may be more difficult
for an algorithm to recognise and this may contribute to different quantitative recorded data.

**Language detection algorithm background**

The study then explored the language detection algorithm using the same approach, examining the 56 articles retrieved as archival evidence in this chapter as a starting point to which supplementary articles identified using Google as a search tool for the terms “language detection” + “algorithm” + “Facebook” were added. The data documented concerns the ability of language detection algorithms to cope with vast amounts of harassment and hate speech on the platform, which required artificial intelligence systems that could understand the shifting nuances of more than 100 different languages (Simonite, 2018). Any shortfalls were required to be identified by Facebook’s backup filtration staff of 15,000 human reviewers (Fick & Dave, 2019). Considering the amount of content required to be reviewed at scale, the workload is unmanageable for these 15,000 employees as documented by in a report by Barrett (2020). Because of staff costs, there are questions about the sustainability of this process, and it highlights that the current management process may be under-resourced and is destined to put a strain on the service, and pressure on its employees. Gaps in the enforcement net can have harmful real-life implications for victims, as discussed in the 4.2 impact findings (on p. 64) of this study. Many severe trolling examples in the Australian news media have been documented to inflict long-term damage, and thus emphasise the need for social media organisations, including Facebook, to resolve funding and business model issues to ensure the safety and protection of its users. Although Microsoft and Alibaba claimed on January 15, 2018, that its software could read like a human (He et al, 2015), software’s potential to deduce content meaning is still a work-in-progress. The continuous language progression of computers has enabled virtual assistants such as Alexa significant advancements in the accuracy of automatic translations. However, critics still argue artificial intelligence is still a long way from understanding even relatively simple text in the way humans do (Simonite, 2018, para. 13). The following statistics were retrieved from Facebook’s (2018) community standards
enforcement report to analyse and interpret the effectiveness of the language detection algorithm. This study identified and calculated:

- In one year, 2.1 million posts were flagged to contain bullying and harassment, and were subsequently removed from the platform, between the October 1, 2017 and September 30, 2018.
- For the same time-period, the language detection algorithm had an average success rate 14.9 percent for identifying bullying and harassment violations before being reported by users.
- This meant the algorithm failed to detect 1.79 million bullying and harassment violations during that year.
- The study further identified the algorithm failed to identify any bullying and harassment violations before being reported by users until July 2018.

For comparative illustration, the study further identified and calculated:

- In one year, 8.6 million posts were flagged to contain hate speech and were subsequently removed from the platform, between October 1, 2017, and September 30, 2018.
- For the same time-period, the language detection algorithm had an average success rate of 41.5 percent for identifying hate speech violations before being reported by users.
- This meant the algorithm failed to detect 5.03 million hate speech violations during that year.

**Language detection algorithm thematic analysis**

Through thematic analysis of qualitative and quantitative data results of language detection algorithms that help to protect Australian journalists from being trolled, one central theme emerged: decoding problems. Limitations of the software are revealed in the translation of language, particularly when using slang terminology. Foer states that “algorithms can translate languages without understanding words, simply by uncovering the patterns that undergird the construction of sentences” (Foer, 2017, para. 25). However, the language detection algorithm is far more complex, and is required to interpret language, rather than translate.
Facebook’s Chief Global Security Officer Nick Lovrien (cited in Bell, 2018) explains the shortfall of the system is the inability to interpret human interactions based on idiomatic expressions and suggests regulating the intricacies of languages requires human review. Native language of a metaphorical nature can be misinterpreted through the translation process, if read in a literal context. Lovrien (cited in Bell, 2018) further explains saying, “in English we say it’s ‘raining cats and dogs’ … but in Portuguese, it’s ‘raining knives and daggers” (para. 10). A Facebook algorithm can identify content that contains the word “knives” (para. 10) but, without context, does not have the ability to determine if the comment is violent, or talking about the weather. The English language is particularly hard to decode, as the underlying meanings of many expressions are not represented in the literal text (Knight, 2016).

The impacts of trolling are subjective by nature, and victims are impacted differently by certain content. Offense is not universal and, therefore, a universal algorithm is destined to fail. As a notable limitation, language detection algorithms are limited in their ability to understand context and, therefore, cannot always determine the intent of a comment. Problems arise with comments such as “I’m going to beat you,” which could be a threat or friendly competitive challenge. Because of this complexity, there is no simple techno fix to the problem and people need to be engaged with finding a solution. Previous research on social problems has found that the people who need to be engaged are not just the victims and perpetrators, but stakeholders on all levels of the systems model (Rittel, 1972, Funnel, 2014; Beutler, 2021). In a 2021 study, Australian researchers Aim Sinpeng, Fiona Martin, Katharine Gelber and Kirril Shields called on Facebook to boost its consultation with minority groups and better train its in-house moderators in response to online hate speech in the Asia Pacific region. The was the first Facebook-funded regional study and found despite automated moderation and improved content standards, the platform is still allowing vilification and discrimination to proliferate on public pages, with inadequate mechanisms to address it (Sinpeng et al, 2021).
Trolling content depicted in image often contained nudity and violence that were more universally recognised as offensive and Facebook algorithms were thus more effective at identifying trolling delivered through a visual medium. However, additional decoding flaws were revealed by which trolls could avoid language detection algorithms, hiding abusive text within images and videos. To combat this, Facebook announced a new program in 2018 called Rosetta that was designed to read text and divert flagged content for moderation. Academic discussions recognised decoding problems as “important”, yet progression as “unsatisfying” (Huang, cited Simonite, 2018, para. 14) with Huang further concluding that “current models are not that intelligent, that’s the problem.” Srinivas Narayanan, who leads engineering in Facebook’s Applied Machine Learning group, agreed with frustrations with the algorithms’ current limitation of decoding language. Although progress in image detection enables the system to scan for pornography and hate speech on a vast scale, Narayanan (2018) suggested human-level accuracy and nuance remains a distant hope for language detection. Although Narayanan (2018) believed with time artificial intelligence could eventually achieve this desired level of interpretation, he stated “we just don’t know how yet” (Narayanan, cited Simonite, 2018, para. 15). Facebook continues to employ a large, multinational team to conduct long-term fundamental research with aims, one day, to bridge the gap between autonomous recognition and understanding.

**Practical application of machine learning and artificial intelligence**

**Functional strategy quantitative data results**

It was important to at least consider the practical application of machine learning and artificial intelligence models in reducing trolling on Facebook, although a comprehensive analysis to determine their effectiveness fell beyond the scope of this study. However, for a general discussion of their influence, the study explored Facebook’s website https://transparency.fb.com and retrieved the Facebook Community Standards Enforcement Data Snapshot (2018a) displayed in Figure 28 (on p. 191). As Facebook integrates both image detection and language detection algorithms together, the retrieved data assessed their combined effect on identifying abusive content. Thematic analysis applied to the aggregated data
revealed bullying and harassment was the most challenging problem for the artificial intelligence software to identify, followed by hate speech. Facebook acknowledges this in its corporate communications. These two themes are highlighted in red in Figure 28 for visual comparison.

The data indicates artificial intelligence algorithms do not work autonomously in response more complex and subjective concepts, and subsequently fail as a comprehensive proactive combative feature against trolls.
Chapter 4.5 Facebook response results and discussion

Note: The image is produced by Facebook in 2018. Bullying and harassment and hate speech are identified as difficult areas for automatic detection.
Qualitative thematic analysis

Through thematic analysis of the Facebook Community Standards Enforcement Data Snapshot (Facebook, 2018a), six themes emerged: the vastness of the content; impersonal automation; user uptake; lasting harm; censorship bias and fake news. Each theme is discussed and critiqued in the following sections.

Vast content

With over 2.85 billion users at the first quarter of 2021 publishing over 2 billion posts every day (Statista, 2021), Facebook’s vast content archive has burgeoned to unmanageable levels that require technological intervention. Critics have suggested that Facebook was reluctant to commence research and funding to create artificial intelligence feature responses due to “ideological” hopes (Taibbi, 2017, para. 12) that the trolling problem would subside on its own over time. However, trolling frequencies have not subsided and the dilemma of monitoring content at scale remains. Critics, including Taibbi (2017), further suggested that prior to 2016, Facebook had a logistical nightmare because “keeping porn and beheading videos out of users’ news feeds was an extraordinarily involved technical process” (para. 13). Examples of extreme violence initially took precedence over the impact of trolling on victims, and resources were disproportionately distributed towards image detection algorithms rather than language detection models. In a multi-layered approach mirroring Ackoff’s (1974) framework that this study adopts, Facebook used a combination of reports from users to which artificial intelligence models were applied, to more quickly identify content that violated their Community Standards. Flagged content was reviewed by Facebook’s Community Operations team who worked on rosters operating twenty-four hours a day, seven days a week, in over 40 languages (Facebook, 2018c). In 2018, Facebook had more than 15,000 content reviewers, an increase of over 40% from the previous year (Facebook, 2018c). This marked the most responsive time period by Facebook in implementing techniques to combat trolling on the platform. This coincided with the saturation period identified in chapter 4.1 (on p. 62). While the improvement in moderation by Facebook is likely to have reduced the risk that journalists would be exposed to trolling, given the vastness of the platform, gaps likely remained.

Chapter 4.5 Facebook responses results and discussion
**Impersonal automation**

Journalist users have expressed frustration at automated responses and criticised computing algorithms for their impersonal nature (Munn, 2020). The study identified the importance of an easy-to-use reporting process to moderate the described frustration levels and encourage trolled journalists to use the report feature. Simply removing the offensive images without acknowledging the trauma that targeted victims and moderators endured was highlighted as a concern that Facebook’s research team was deliberating in 2019 (Facebook, 2019). This is relevant to the thesis because the automated filters used by Facebook and its reporting mechanisms are created to deal with vast content and are not able to be person or content specific. This enables opportunities for cunningly worded trolling of specific journalists.

**User uptake**

Facebook reported that some of its users who had been subjected to abuse were not familiar with the online reporting process and were unsure how to navigate the website to locate the feature tools. Facebook (2019) reported this lack of knowledge as an identified area of concern that researchers were attempting to resolve in 2019. This indicates the importance of educating journalists and comment moderators in social media platform capabilities and reporting processes.

**Lasting harm**

The study identified that victims whose images were shared, or who were threatened online, felt violated, angry and embarrassed and were often “scared and worried that their family, friends, and co-workers would see the images” (Facebook, 2019, para. 4). The study revealed that harm continued for victims long after the images were removed. Facebook (2019) reported mental health consequences that included anxiety, depression, suicidal thoughts, and sometimes post-traumatic stress disorder (para. 4) and that further highlighted the potential economic and professional consequences for victims, including lost jobs, fewer professional connections, and colleagues who tease or avoid them. Facebook further reported difficulties for victims to find new employment and suggested, in
many cases, the costs to victims are “serious” and “unquestionably long-lasting” (para. 4), this impacts the career trajectories of trolled journalists.

Censorship bias

Facebook’s Community Standards project puts the company in the position of deciding arbitrarily what speech is acceptable and what is not which is then enforced by the algorithms. Tepper and Hearn (2019) outlined their perspective stating “we may fool ourselves into thinking that Facebook and Google use fair, impersonal algorithms to monitor speech. But algorithms are programmed by people, and people are imperfect and have biases. The left may be happy that conservative biases are censored today, but who will control these platforms in five to 10 years” (p. 92). The effectiveness of algorithms to remove certain content may not only benefit trolling victims including journalists but may be leveraged as a means of societal control.

Fake news focus re-emerging

Through the thematic analysis of archival evidence about the technological developments of Facebook’s artificial intelligence, this study identified an initial shift in focus on what the company had prioritised as the most important factor to address. Although the autonomous removal of abusive content directed at its users on the service was prominent around 2010, the company's main objectives began to digress from harassment as a primary focus, prioritising attention towards the regulation of fake news on the platform around 2016. A UNESCO study in 2018 outlined many uses of term ‘fake news’ to describe “reporting with which the claimant does not agree” (Ireton & Posetti, 2018, p. 45). The study provided the following context to more accurately represent the meaning of ‘fake news’ in reference to trolling and its implications for journalists:

Unfortunately, the phrase is inherently vulnerable to being politicised and deployed as a weapon against the news industry, as a way of undermining reporting that people in power do not like... it covers satire and parody, click-bait headlines, and the misleading use of captions, visuals or statistics, as well as the genuine content that is shared out of context, imposter content (when a journalist's name or a newsroom logo is used by people with no connections to them), and manipulated and fabricated content.
Ireton and Posetti (2018) further suggested, “this crisis is much more complex than the term ‘fake news’ suggests. If we want to think about solutions to these types of information polluting our social media streams and stopping them from flowing into traditional media outputs, we need to start thinking about the problem much more carefully” (p. 45). This study suggests the attention shift at the time was significantly impacted by the US Presidential election campaign of Donald Trump that year. As discussed in chapter 4.1’s emergence findings (on p. 39), the presidential campaign is identified as a significant contributing factor that fundamentally shaped the evolution of trolling in the Australian news media. The election marked the commencement of the saturation period in chapter 4.1. At the time, information surrounding trolling was so widespread that its meaning shifted and flooded news media channels with misinformation leading to public confusion and decreased trust in journalism. Although the election occurred overseas, the impact still influenced the Australian news media, as it did around the world. Facebook campaigns against inauthentic content took precedence, which was detrimental for the progression towards any tangible support strategies for trolled victims. Critics further suggestions that after Trump’s presidential win, “everyone turned to Facebook and Google to fix fake news, but nobody had a coherent definition of what constitutes it” (Taibbi, 2017, para. 18). Wardle (2017) does, however, comprehensively explain that fake news is more than just news, suggesting it is an entire information ecosystem of misinformation and disinformation. Political smear tactics transformed the collective understanding of trolling, and hindered attempts to remove it, and subsequently fuelled its further escalation into 2017 when many Australian journalists were forced to leave social media as a result (Ford, 2017; Noyes, 2017).
4.5.4 Facebook research strategies results and discussion

Research strategies were identified by the study as the fourth type of Facebook response to trolling that emerged from analysis of the 56 articles retrieved as archival evidence in this chapter. This study noted a complete assessment of all research initiated by the platform was not feasible. However, one in particular was referenced in four of the 56 articles and was selected for further investigation: Facebook’s non-consensual intimate imagery research. Using Google as a search tool for the terms “Facebook” + “intimate imagery” + “research,” a further 11 relevant articles from credible newsrooms were retrieved that described or made comment on Facebook’s non-consensual intimate imagery research, which are summarised and presented to illuminate common themes for discussion and critique. As the research developed, platform adjustments were progressively implemented throughout 2019 that addressed concerns of gender-bias. The study adopting the feminism paradigm as a framework for analysis is particularly relevant to this section because it is about gender.

To ground the discussion in the context of Australian trolled journalists, the study spotlights the malicious, targeted abuse of female journalists by reiterating chapters’ 4.1 and 4.2 findings that women are disproportionately affected by trolling. Not only are female journalists trolled more frequently, but they are trolled more aggressively than males, with threats of physical violence and sexual abuse a daily occurrence in newsrooms. The explicit content directed at females is echoed in user sentiments on the platform that highlight a gender imbalance. By failing to eliminate trolling, Facebook has continued the facilitation of pathways used by abusers to exploit vulnerabilities, and further deepen the gender divide.

Non-consensual intimate imagery research background

The term non-consensual intimate imagery is defined as a form of sexual violence that can be motivated by an intent to control, shame, humiliate, extort, and terrorise through the act of exposing personal images of woman online (Facebook, 2019a). The term, abbreviated to NCII, is often referred to as revenge porn, although revenge is not always the motivating factor. Consequences of sharing intimate images include grave and lasting emotional and physical impacts for
victims. Between 2018 and 2019, Facebook in partnership with several international safety organisations, conducted research on the impacts of NCII, to review and improve responses to trolling. In a comprehensive approach, research findings were implemented across Facebook, Messenger and Instagram, the applications owned by Facebook Inc, which also owns WhatsApp. Facebook said it aimed to understand experiences through the lens of victims with the objective to better identify barriers to the reporting process that made users feel unsafe on the platform (Iyengar & Nain, 2019). Facebook conducted interviews with victims and support advocates from Kenya, Denmark and the United Kingdom and brought together over 20 academics and non-profit leaders from 10 countries in a conference, to broaden collective understanding of best-practice (Davis, 2019). The conference flagged technology, educational information, and psychosocial support as the most influential factors for future effective proactive responses (Davis, 2019). A combined approach of user precaution (to minimize behaviour susceptible to attack) and refined platform responses (to monitor and restrict communication avenues) were suggested by the study as the best responsive approaches to NCII trolling.

**Optional profile picture guard**
Non-consensual intimate imagery research provided Facebook with a greater understanding of global trends facing women and enabled practical features to be implemented in response to trolling. Cultural discrepancies revealed some Indian women chose not to share profile pictures of their face due to growing concerns that photographs could be exploited by trolls to shame or dishonour victims and their families (Soman, 2017). In response to these concerns, Facebook developed an optional profile picture guard that gave women in particularly vulnerable regions more control over who could download or share their pictures. This feature may have been helpful for journalists such Maria Ressa, with a 2021 study by the International Centre for Journalists finding that 40% of her trolling attacks were personal assaults, with 14% classified as misogynistic, sexist and explicit abuse (Posetti, Maynard, & Bontcheva, 2021). The feature is available in India, Pakistan, Egypt, Bangladesh, and Sri Lanka, where women expressed similar
concerns (Affi, 2018). However, in 2021 this feature is not available in Australia and therefore provides no protection for Australian journalists.

**Image hash strategy**

Facebook introduced a controversial proactive strategy to combat non-consensual intimate image trolling on November 9, 2017. In an attempt to help users who feared that a private photo may be exposed, the platform trialled an unprecedented provocative strategy in Australia (Statt, 2017). Although Facebook developed the strategy “in partnership with international safety organisations, survivors, and victim advocates” (Iyengar & Nain, 2019, para. 9), the public expressed serious concerns with the suggestion that nude photos or videos be directly uploaded to Facebook’s Community Operations team who personally review each image and “hash” it. This process creates a digital footprint that prompts algorithms to block the same image being uploaded elsewhere. Although someone at Facebook is indeed looking at the nude photos, the company stressed that these were “specially trained representatives” (Statt, 2017, para. 4) and dismissed criticisms that the measure was unethical and “counterproductive” (para. 2). Despite the feedback, Facebook claimed its research indicated, “this was an option that victims generally wanted built into the reporting process” (Iyengar and Nain, 2019, para. 10). However, the pilot program would only work if the potential victim was aware that a particular image was at risk of being shared, which is often not the case. The announcement by Head of Facebook Global Security, Antigone Davis, provided the following explanation:

> Australians can complete an online form on the eSafety Commissioner’s official website. To establish which image is of concern, people will be asked to send the image to themselves on Messenger. The eSafety Commissioner’s office notifies us of the submission (via their form). However, they do not have access to the actual image. Once we receive this notification, a specially trained representative from our Community Operations team reviews and hashes the image, which creates a human-unreadable, numerical fingerprint of it. We store the photo hash—not the photo—to prevent someone from uploading the photo in the future. If someone tries to upload the image to our platform, like all photos on Facebook, it is run through a database of these hashes and if it matches, we do not allow it to be posted or shared. Once we hash the photo,
we notify the person who submitted the report via the secure email they provided to the eSafety Commissioner’s office and ask them to delete the photo from the Messenger thread on their device. Once they delete the image from the thread, we will delete the image from our servers.

(Davis, 2017, para. 4)

Uptake of the strategy was relatively low in the first years of operation, which the platform attributed to deficient awareness of the strategy; however, critics suggested a lack of trust by the public in Facebook’s ability to safely store such personal images (Martin, 2017a). Facebook did not concede that the provocative design was more likely a contributing factor, with many victims understandably concerned about sending private intimate images to strangers for external mediated review. From a victim perspective, the strategy is counterintuitive. This study suggests the controversial nature of the proactive strategy represents the dominance of the patriarchal-perspective, and when scrutinised through an alternate paradigm, it can be considered culturally insensitive and potentially offensive in its very suggestions (Soman, 2017). Sexual imagery is forbidden by some religious ideologies, and to request that victims send such photos to their government or a similarly authoritative organisation is highly disrespectful and fails to acknowledge the risks associated with doing so. Although this study acknowledges the shortcoming of prior strategies and the need for new and innovative resolution methods, the practical consideration of this strategy is deemed to be questionable.

**Qualitative thematic analysis**

From thematic analysis of Facebook’s non-consensual intimate imagery research, five themes emerged: deflection of responsibility; humiliation of women; cultural differences; cultural insensitivity; and a holistic approach. All these themes are discussed and critiqued in the following sections.

**Deflection of responsibility**

Facebook often seeks assistance from affiliates overseas. In Pakistan, Executive Director of the Digital Rights Foundation, Nighat Dad, has worked with Facebook as part of the management team reviewing harassment issues on the service, and
in particular, the sharing of non-consensual intimate images. She stated, “online gender-based violence is not a technology problem, it is a societal problem” (Dad, cited in Facebook, 2019a, para. 21). Although the fundamental argument is correct, because of the scale of social media platforms, trolling has emerged as a technological problem as well. The Facebook management teams and associated management affiliation are often quick to deflect blame, which is further discussed in the next theme below. Although trolling is not perpetrated by the social media platforms themselves, their creators have provided the forum from which trolls exert their abuse. Public criticism asserts that partial responsibility for its prevention falls on the facilitators (Leetaru, 2018; Morell, 2021). Prior to 2014, identified as the social problem period in this study (and described on p. 55), when the recognition of trolling as a standalone social problem was not universal, Facebook founder Mark Zuckerberg shunned this responsibility (Taibbi, 2017). Indeed, Zuckerberg deflected accountability as late as June 2016, saying, “editing content... that’s not us” (para. 11). Although Facebook does not make journalism, it is hosting, distributing, and monetising content just like a media company, and criticisms suggested a declaration of partial responsibility should have been made (D’Onfro, 2016). Continued deflection of responsibility halted any progress towards the solution of the trolling problem for several critical years. As trolling evolved and permeated across the service, Facebook had no alternative than to commence more in-depth counteractive strategies.

**Humiliation of women**

Facebook’s research responds to the trend of trolls attempting to humiliate women as one of their primary tactics. Examples of revenge porn involving female journalists include British journalist Anna Richardson in 2015 (Horswill, 2015), Scottish journalist Vonny Moyes in 2016 (Ballard, 2016), and Australian journalist Fiona Harper in 2019 (Wondracz, 2019). Abusers can also obtain or create via Photoshop private images of women, often nude, posing in suggestive positions or engaging in sexual actions, and share these images without consent on Facebook to publicly embarrass and discredit the victim. Although Facebook claims to remove such images under their standards on the sexual exploitation of adults, with 1,400 cases of revenge porn abuse reported across Australia in 2019 (Taylor
& Mitchell, 2019), examples continue to materialise and circulate undetected suggesting Facebook’s removal procedures are flawed and inconsistent. This particular tactic crosses legal boundaries discussed in further detail in the 4.7 legal responses section of this study (on p. 265).

**Cultural differences**

The internet has both facilitated globalisation and fuelled cultural interactions previously constrained by the cost of international travel. Through the introduction of the internet, cultures were able to spread exponentially, with different beliefs, practices, and ways of living able to be explored and experienced by anyone with access to a computer. Online content was shaped by cultural diversity (Burri, 2016) but was not immune from being exploited for targeted abuse. Many Silicon Valley technologies were encoded with Western biases that prevented users from learning more deeply about other people, cultures, and places (Srinivasan, 2017). To support richer understandings of differences, social media networks need to be redesigned to better represent diverse cultural and political perspectives and reflect a more inclusive approach when determining which content is offensive and to be removed. Different cultural backgrounds have varying tolerances of certain behaviours and although it would be considered shameful by the broader Australian community for a woman to have taken photos while performing a sexual act, in more traditional communities, women may face extreme repercussions and be harmed, shunned, or excluded from their communities (Facebook, 2019a). While bikinis are commonly worn in Australia, a woman Saudi Arabia would be shamed if a troll revealed a photo of her ankle and put at risk of physical harm if a troll published a photo of her walking with a man who was not an immediate family member (Facebook, 2019a). Facebook highlights that many women were forced to run away from home to avoid persecution after imagery was published on the platform. When considered in its full cultural context, images shared in this way are intentional attempts to humiliate women. In 2021, revenge porn had become a political weapon in Myanmar with journalists such as Han Nyein Oo profusely targeted (Thiha, 2021). Facebook recognises cultural discrepancies and is required to monitor trolling content based on its individual merits, to ensure examples such as this are specifically removed under Facebook’s
Cultural insensitivity

The exponential rise of popularity, usership and the fundamental integration of Facebook within contemporary culture can be considered a natural progression of technological advancements. However, this rapid success was not a coincidence but an intentional objective at the foundation of Facebook’s construction. As spotlighted by Nussbaum (2010), Facebook purposively matched its technological design with a deep Western cultural longing characterised by the adolescent desire for connection. With this theoretical perspective ingrained, response strategies were shaped by its cultural biases. In a 2018 study, researchers Hong and Na (2018) concluded, “cultural backgrounds determine not only the features of online networks, but also the patterns of online behaviours” (p. 441). Therefore, because of complex cultural nuances, universal strategies deployed at global scale were likely to fail, with non-western cultures often the first casualties. While, this study recognises that distinctions between western countries also exist, and current universal approaches to trolling may need to be adapted to each geopolitical region. Facebook is used in different ways in independent and interdependent cultures (Hong & Na, 2018). Individualism, the dominant western perspective, governs much of cyber-hate literature (Jane, 2016), in a self-focused approach. Collectivist perspectives favoured in interdependent cultures, including China and Korea, focus on belonging and social obligation (Ren, Cang & Ryder, 2021). The study does not infer any conclusions about differences in trolling frequency or severity between idiosyncratic cultural characteristics; however, it does suggest that the current individualist, self-focused, western approach has not eliminated the problem to date, that a more holistic multi-faceted social approach to trolling responses, sensitive to a wider array of cultural principles, may work better.
Holistic approach

Describing Systems Theory, Ackoff (1974) wrote, “the properties or behaviour of each element of the set [or system] has an effect on the properties or behaviour of the set taken as a whole” (p. 13). He went on to explain that this interconnectedness is why a system’s performance depends on how well the parts fit and work together, not just how well each one performs individually. This thinking seems to have permeated Facebook’s research strategy which acknowledges the importance of a holistic approach to the complex social problem of trolling stating, “we take a comprehensive approach to making our platform a safer place for women, including writing clear policies and developing cutting-edge technology to help prevent abuse from happening in the first place” (Facebook, 2019a, para. 5). Facebook reiterates the significance of a coordinated and multifaceted approach to trolling on its platform stating, “to account for this wide spectrum of harassment types, our rules need to be thoughtful and similarly comprehensive” (para. 11) and “blocking and reporting are only part of the solution, and their success relies on people knowing to seek them out and understanding how to use them, plus feeling comfortable enough to use them” (para. 14). Although Facebook repeatedly invokes the need for a comprehensive approach, it fails to effectively outline the required processes for the successful implementation of this strategy. Facebook does provide data in some cases, however, on a global/Australian level it is lacking. The current toxic environment highlighted by recent examples of Australian trolled journalists (Hoffman, 2020; Smith, 2020; O’Callaghan, 2021) illustrates its failed implementation at present.

4.5.5 Facebook corporate approach strategies

The notion of corporate social responsibility was not new when Ackoff formulated his model in 1974. He wrote:

Mounting public pressure on corporations to behave in more socially responsible ways has given rise to a major debate in corporate circles. The debate focuses on the question: Should or should not a corporation involve itself in activities and programs that are intended to improve its physical and social environment? This question arises only when such involvement consumes corporate profits; when it produces profit there is no issue. The answers to this question reflect the four basic attitudes towards the present
and future; inactivism, reactivism, preactivism, and interactivism.

(p. 56).

He described inactivism, reactivism, and preactivism as profit-focussed, while citing R.N. Anthony's indictment in the *Harvard Business Review* of profit as immoral and socially unacceptable. Although Ackoff conceded that profit was also the most universally accepted corporate objective. He then deployed the argument that survival is the central purpose of a company and while profit may serve that purpose, social responsibility also plays a role. Making this point, Ansoff (1965) had said firms should aim to balance the conflicting claims of various stakeholders (managers, workers, stockholders, suppliers, vendors) and to configure its objectives to give a measure of satisfaction to each. While this style of management has evolved since the 1960s, the central tenant aligns with what Facebook appears to be doing.

Corporate approaches were identified as the fifth type of Facebook response to trolling. Through exploration of archives on Facebook’s website https://about.fb.com, and analysis of the 56 articles retrieved as archival evidence in this chapter, this study identified multiple corporate responses by the platform. However, in consideration of the scope of the study, the three strategies considered most relevant to Australian trolled journalists were selected for further inquiry and are listed below:

1. **Transparency**
2. **Live-stream continuation**
3. **Moderator compensation**

**Background to transparency corporate approach strategy**

Comparing the first management responses to trolling by Facebook and Twitter (as identified in chapter 4.4), the study revealed a significant difference in levels of concern. While Twitter grappled to admit trolling was a problem, Facebook led by example adopting a transparent corporate approach. Twitter’s first response came on September 14, 2012, in a statement that acknowledged trolling occurred
on its platform and outlined an updated user policy (Twitter, 2012). The 16-word addition to the previous user policy outlined abuse would not be tolerated, without providing any function for users to report such behaviour at the time. In stark comparison, Facebook had implemented a report feature as part of its initial setup and, on June 20, 2012, released the Reporting Guide (Facebook, 2012) that detailed the back-of-house reporting process in an effort to be transparent with its users. Facebook accepted some responsibility for the negative behavioural patterns that had emerged on its platform and adopted a more open and honest approach with the public than Twitter. Facebook management continued to share difficulties, discuss limitations and setbacks, reach out to researchers and partnering companies for specialised support, and attempted to educate users about procedural updates as they developed. The processes illustrated on Facebook’s Reporting Guide in 2012 still reflect its current procedures and provide useful information for Australian journalists who are trolled at work. The infographic is displayed in Figure 29 (Angotti, 2012) on the following page and provides both qualitative and quantitative data for thematic analysis and critique.
Figure 29: An internal progression flow-chart of a Facebook report

What Happens When You Report Something?

At Facebook, nothing is more important than the safety and security of the people who use our service. With a community of over 901 million people, Facebook maintains a robust reporting infrastructure made up of dedicated teams all over the world and innovative technology systems.

In March 2012, Facebook pioneered a tool called 'Social Reporting', which enables people to report problematic content not only to Facebook, but also directly to their friends to help resolve conflicts.

Additionally, people can use the tool to reach out to a trusted friend who may understand the offline context of the situation and are able to assist.

The image is produced by Facebook in 2012 (Angotti, 2012)
Transparency corporate approach strategy thematic analysis

Through thematic analysis of data presented in Facebook’s Reporting Guide from 2012, the three themes that emerged (nonlinearity, systems theory and hierarchy) are unpacked and discussed in the following sections.

Nonlinearity

Social problems often arise from a series of cumulative events as supported by chapter 4.1 findings that mapped the emergence of trolling in the Australian news media. Behaviour dynamics can be analysed for patterns that provide an insight into the underlying structure of a system, key to understanding what is happening and, importantly, why. Social media companies often respond to individual events rather than considering and addressing more complex underlying causes (Gillespie, 2018; Napoli, 2019). Meadows (2008) explains that in approaching complex problems, humans have linear minds in a nonlinear world and suggests “we are often insufficiently skilled in understanding the nature of relationships” (p. 91). At first glance at Facebook’s procedures in Figure 29 it is evident the platform has taken a nonlinear approach, with the flow-chart provides a number of different pathways to report abuse. Linear relationships have an important modular feature that enable them to be dissected and rebuilt multiple times. However, the unpredictability of nonlinear relationships creates “rich kinds of behaviour” (p. 23) that generally do not have a resolute solution, a phenomenon described in the seminal text Chaos: Making a New Science by James Gleick (1988).

Society observes the world in terms of constant proportions only provided by linear models, where cause and effect are relative. However, nonlinear relationships do not produce proportional effects and are therefore more difficult to comprehend and forecast. Nonlinearities confound our expectations of responsive action engendering behavioural change. Facebook’s diverse response web is a constructive approach; however, its effectiveness is difficult to assess in isolation.

Systems theory

Facebook is an example of a complex system, characterised by a large network consisting of relatively simple components working in cohesion. One year after
the launch of Facebook, Skyttner (2005) described six system characteristics determined as applicable to the platform which included: a large number of elements; many interactions among the elements; attributes of the elements not being predetermined; interaction between elements being loosely organised; the system being subject to behavioural influences; and the system being largely open to the environment. The relationships between system components shape the properties of the system more than the nature of individual components. The Reporting Guide appears to implement this theoretical approach to address trolling by focussing on the interconnected components of the reporting system. Although the systems approach is often perceived as an approach for solving technical problems, it has emerged as an adaptable concept that is applicable to both material and nonmaterial phenomena. The application of systems theory approach to the fast growing and complex field of social media was forecast by Darena (2011) ten years ago as an important problem-solving approach to a domain that operates as an interdisciplinary system with tendencies of specialisation. However, the results that stem from its apparent applications by social media platforms are varied. Recognised as an effective response to highly specialised branches of science, the systems approach breaks down components of complex problems enabling communication among specialisations to facilitate combined problem-solving methods. The study suggests continued application of this approach by all stakeholder groups is imperative to more effectively address online trolling behaviours.

**Lack of hierarchy**

Influential systems theorist, Donella Meadows (2008) states systems work with greater efficiency, accuracy, and positive structure when resilience, self-organisation and hierarchy are present. Resilience is defined as “a measure of a system’s ability to survive and persist within a variable environment” (Meadows, 2008, p. 76). Self-organisation means “changing any aspect of a system lower on this list by adding completely new physical structures, adding new balancing or reinforcing loops, or new rules” (Meadows, 2008, p. 159). Hierarchy is defined as “the ordered nature of a system where individuals, groups, or entities are conceptually situated relative to one another based on power or rank” (Mangal,
2013, p. 119) and, in a highly functional system, the welfare, freedoms and responsibilities of each part of the system must be balanced (Meadows, 2008, p. 85). Although Facebook’s Reporting Guide exhibited resilience and self-organisation characteristics, a defined hierarchy was lacking. In support of this finding, Mangal (2013) stated Facebook had “fairly present resilience, very present self-organisation, but minutely present hierarchy” (p. 118). Yet despite hierarchal discrepancies that are still present in 2021, the reporting feature seemed to function well. Although minor hierarchies may exist within parts of Facebook, including the elevated status of Groups and Pages, there is no built-in hierarchy for the entire site and implementation at a user-level may disrupt an interactive function if friends are ranked, a discriminative function of MySpace that contributed to its demise. These junctions could play a part in regulating behaviour on news-related pages if blocking, suspension, or demotion in ranking was a negative consequence of bad behaviour such as trolling. The capacity to report could help trolled journalists overcome the sense of powerlessness they felt. However, the study observes alternate models have successfully contributed to the improvement of behavioural standards through self-regulation: Uber, for example, ranks its users by five stars (Scheiber, 2017), as does eBay and Facebook Marketplace. Such a mechanism could be taken into consideration when adopting preventative strategies to trolling. But while the risk of poor ratings could deter trolls or trolls could use the power to demote people to harass victims, if not implemented at the user-level, it could be more forcefully applied at the corporate level to shift the power imbalance and more efficiently remove the authority of abusers to use the platform. Although some ranking of comments (sorted into top comments and most relevant) does occur, Facebook uses algorithms that are not publicly disclosed to prioritise content. While Facebook has considered internal levels, the platform is part of a larger system and is resistant to being legislated, deflecting legal responsibilities to the publications to shoulder themselves.

**Live stream continuation corporate approach strategy background**

In the analysis of the 56 articles retrieved as archival evidence in this chapter, Facebook’s live stream function was selected for inquiry and collective descriptions were summarised to provide a brief background. Launched on
August 5, 2015, the live stream feature on Facebook evolved from being exclusively used by celebrities with verified accounts, to streaming violent criminal acts. Although majority of viewers accrue after filming has ceased, Hern (2017) suggested the difference from pre-recorded videos was evolutionary, not revolutionary. The live or pre-recorded function of content is not the main concern, but the ease with which unmonitored content can be posted. In 2019, Facebook announced it would hire 3,000 people to “speed up the removal of videos showing murder, suicide and other violent acts” (Tsukayana, 2017). Despite criticisms (Crossley, 2017), exploration of Facebook’s website revealed the feature remained active as of June 2021. Acts of hate-speech and terrorism have and continue to present significant problems for investigative journalists including Finnish journalist Jessikka Aro (Miller, 2016) and Indian journalist Rana Ayyub (Fitzgerald, 2022).

**Key event: New Zealand massacre**

Despite investing in artificial intelligence to aid its 15,000 content moderators in 2019 (Newton, 2019), Facebook’s policing measures failed to flag a massacre in New Zealand that was live streamed on its platform for at least 15 minutes (O'Sullivan, 2019). A central intention of perpetrators publishing violent acts online is to incite a reaction and considering the majority of users would respond in disgust and that the trolls own extremist community may reward the poster for this behaviour, the act can be classified as a form of social trolling, aimed to provoke the broader community rather than to target an individual.

**Live stream continuation corporate approach strategy thematic analysis**

In thematic analysis, the two themes of cumulative scrutiny and anti-racism social movement emerged. The descriptive commentary below provides context for a comprehensive analysis of the implications for Australian journalists.

**Cumulative scrutiny**

In a comprehensive investigation led by *The Guardian’s* Samuel Gibbs (2017), Facebook came under extensive public scrutiny in mainstream media. The following descriptions of key events formed the basis of this criticism. On April 26,
2017, a man live streamed the murder of his 11-month-old daughter in Thailand (Gibbs, 2017). Relatives reportedly saw the distressing footage and alerted the police who arrived too late to save the 11-month-old or the father, who had committed suicide. The two posted videos were accessible to Facebook users for approximately 24 hours before being taken down and had been viewed 112,000 and 258,000 times respectively. The videos were uploaded by other people to YouTube but they were removed within 15 minutes of YouTube being notified. This study identified the murder as a catalyst event that played a pivotal role in shaping public perception of Facebook’s inaction against online extremism. It was not an isolated example, and a cluster of extreme acts of violence broadcast on the platform drew attention to Facebook’s inaction and widespread condemnation of it at the time. Additional incidents include the torture of an 18-year-old disabled man on January 6, 2017 that was live streamed and watched by more than 16,000 users (Levin, 2017) a video which is concerningly still accessible online as of July 2020; the sexual assault of a 15-year-old girl by six men on March 22, 2017 live streamed to more than 40 users (Associated Press, 2017); the random shooting of a 74-year-old retiree collecting aluminium cans on April 17, 2017; and although not broadcast live, video of the murder was posted on the platform shortly afterwards and remained visible for over three hours (The Guardian, 2017). The ease with which the platform can be used as a podium for hatred and violence incited public anger and outrage, yet Facebook still continued the feature despite a storm of scrutiny (Gibbs, 2017).

Precursor for the Black Lives Matter movement

Intending to increase interactivity and connectivity amongst users through a live streaming function, Facebook may not have predicted the cultural impact of the feature. In essence, the function shifted the power balance in favour of users and negated traditional gatekeeper roles by facilitating the dissemination of unmonitored content in real-time, enabling users to harness the puissant tool and act as citizen journalists, and citizen pornographers, and citizen executioners. Live streaming provided unedited documentation of police hostilities that became catalysts for countering activism. The first death live streamed on Facebook occurred on July 6, 2016, when Diamond Reynolds recorded the immediate
aftermath of a police officer shooting her boyfriend, Philando Castile. The video shows that after being asked to show his license and registration, Castile advised in a calm demeanour that he was legally carrying a gun before reaching for his license and being shot seven times. Castile’s was the first of several deaths in police custody live streamed on Facebook. Disparate discourses, once considered anecdotal evidence, began to form a cumulative record that exposed the existence and extent of the police racism and discrimination, and through this awareness, public perceptions were shaped. The association between the live stream function and police killings was the precursor for the Black Lives Matter movement. In support of this idea, Hern (2017) suggests this association is about more than just being the only camera available. The difficulty of removing online content has emerged as a characteristic of importance when dealing with an untrusted adversary with the legal power to confiscate devices and suppress footage. The immediacy of live streaming can alert personal contacts and increase the number of people who witness an event beyond those physically present and be further leveraged through increased exposure as a call to arms for the fight for justice. In a snowballing effect, Facebook algorithms aggressively push live streamed videos to the top of newsfeeds in an effort to encourage the interaction; however, this promotional feature also accelerates the adverse rapid exposure of offensive content to large audiences. As acknowledged by Facebook (Rosen, 2019) and journalists (Cox, 2019; Silva, 2021), when abuse happens in real-time, it is more difficult for social media platforms to control as the standard reporting method that requires external moderation is delayed.

**Catalyst for the Black Lives Matter movement**

This study identified the Facebook live streaming feature as the catalyst for the Black Lives Matter movement, marked by its global peak on March 25, 2020, broadcasting on Facebook the death of George Floyd in police custody. The movement called for change worldwide. The need for change was evident in the racist content reported by Australian journalists described chapter 4.3 (on p. 82).
Absence of dump button

While traditional television broadcasters are subject to regulations, internet streaming services do not face the same restrictions and are left by government bodies to regulate and enforce their own terms of service. With a sole reliance on reactive reporting methods, through a combination of people-powered and automated means, the challenge for companies that host videos is to discern objectionable content in real time, a process that none of the large social media platforms seems capable of doing. Facebook’s lack of responsiveness on removing shootings, beheadings, gang rape and child sex abuse from the platform raises questions about the reasons for Facebook’s refusal to discontinue live streaming from its service. Consideration of a delayed broadcast may provide a buffer for automated algorithms to flag abusive content for removal (Kanter, 2019). The success of radio’s seven-second dump button continues today and may be a possible avenue for further research into live streaming. This function would still allow the desired real-time interactivity between users who respond to comments within their shared videos but may provide greater scope for identifying of violent and offensive imagery. The apparent impunity with which the platform can be used as a podium for hatred and violence also raises concerns about employee welfare of those tasked to examine and filter the obscene content (Newton, 2019).

Moderator compensation corporate approach strategy background

In the 56 articles retrieved as archival evidence in this chapter, discussions, or concerns about moderator compensation on Facebook were raised in six articles. Although the scope of the study does not permit an in-depth exploration of adequate moderator compensation, collective descriptions were summarised to provide a brief background of key concepts for further discussion and critique.

Cumulative concerns for Facebook moderators

On May 25, 2017, The Guardian spotlighted growing psychological concerns in the article, “Underpaid and overburdened: the life of a Facebook moderator”. Testimonies from those working to keep beheadings, bestiality and child sexual abuse images off Facebook indicated support procedures were commonly insufficient, and in some instances non-existent. Interviews with staff reflected
immense workplace distress. One stated, “there was literally nothing enjoyable about the job, you would go into work at 9am every morning, turn on your computer and watch someone have their head cut off. Every day, every minute, that is what you would see, heads being cut off” (Solon, 2017, para. 1). The respondent earned roughly $15 per hour removing terrorist content from the social network after a two-week training course. In addition to training and support reported as “insufficient” (para. 7), Facebook staff reported severe impacts, illustrated in the comment: “people would have to visit psychologists every day as they could not sleep or had nightmares” (para. 5).

In comparison, the study assessed the preventative workplace procedures of the Internet Watch Foundation (IWF) in the United Kingdom (Solon, 2017). Before a moderator is hired, the applicants are assessed for suitability by a psychologist, who asks about their opinions on pornography in general, their support network, childhood, and triggers. Once they pass this stage, candidates are interviewed about their work skills before moving onto the final stage designed to see how the potential employee copes and letting them decide whether they wish to continue with the role. At this stage, alongside two trained IWF employees, candidates are exposed to progressively more egregious child sexual abuse imagery that concludes with the worst kinds of sexual violence against children. If the job is accepted, moderators have an enhanced background check before they start six months’ training, which involves understanding criminal law, learning about the dark web and, crucially, building resilience to looking at traumatic content. This six-month training period provided by the IWF far exceeds Facebook’s two-week training course.

On September 21, 2018, Selena Scola, a former Facebook content moderator who worked on contract for nine months, filed a lawsuit in the San Mateo County Superior Court that claimed exposure to violent images caused her post-traumatic stress disorder (Garcia, 2018). Scola argued the company failed to protect her and other contractors as they viewed distressing videos and photographs of rapes, suicides, beheadings, and other killings. The unprecedented case was the first to hold Facebook accountable for failing to protect its employees from trolling in the

Chapter 4.5 Facebook responses results and discussion
workplace. The lawsuit stated 7,500 moderators around the world sift through 10 million potentially rule-breaking posts every week. Facebook relies on its two billion users to report inappropriate content. The moderators then employ the hundreds of rules Facebook has developed to determine if the content violates its policies. Scola urged Facebook to establish a fund to create a testing and treatment program through which current and former content moderators, including moderators employed by a third party, can receive medical testing and monitoring including psychiatric treatment (Garcia, 2018).

**Pay rise for Facebook moderators**

On February 15, 2019, The Verge journalist Casey Newton published *The Trauma Floor* a scathing investigative report of the secret lives of Facebook moderators and the horrendous content and workplace conditions they endure. The report revealed Facebook contractors in Phoenix were suffering from long-term mental health issues after working as content moderators, as their jobs required them to view a steady stream of violent and disturbing content, resulting in many struggling with PTSD-like symptoms (Newton, 2019a). In response to public criticism, on May 13, 2019, Facebook announced a pay increase for its content moderators, up $3 US per hour to $18 US per hour, (Newton, 2019a).

**Mandatory counselling support for Facebook moderators**

In further response to the growing negative discussions that preceded *The Verge* report, Facebook stated it would now provide access to on-site counselling during all hours of operation for its workers, rather than only during the day shift, and would also begin surveying contractors about their mental health twice a year and use these results to shape programs and practices (Newton, 2019, para. 4). Moderators would also be able to blur graphic images before viewing them, to reduce the risk of being caught unawares by disturbing content. However, as a result of accumulating pressure, the implemented strategies were considered by the judge too little too late (Newton, 2019), and Facebook was successfully sued by moderators seeking damages.
Facebook moderators compensated

On May 12, 2020, a landmark acknowledgment of the toll that content moderation takes on employees was made in the case Scola, et al. v Facebook Inc 2020. In a preliminary settlement Facebook paid $52 million US to both current and former content moderators in compensation for mental health issues developed on the job (Newton, 2020). Each moderator received a minimum of $1,000 US and would be eligible for additional compensation of up to $50,000 US if diagnosed with post-traumatic stress disorder or related conditions. The settlement covered 11,250 moderators, and lawyers in the case believed more than half may be eligible for additional compensation related to mental health issues associated with their time working for Facebook, including depression and addiction. Plaintiff lawyer, Steve Williams (cited in Newton, 2020), said the unprecedented case signified “the harm that can be suffered from this work is real and severe” (para. 3). This study identified the announcement to compensate content moderator employees as the most significant social media response. The response acknowledged the impact trolling has on people, beyond the targeted victim.

Moderator compensation corporate approach strategy thematic analysis

From thematic analysis of collective descriptions from six articles that discussed or raised concerns about moderator compensation on Facebook, four themes emerged: accountability, international moderators excluded, resilience and legal precedence. These are explored and unpacked in the following sections.

Accountability

Up until the settlement, Facebook had acknowledged the escalating presence of trolling on its platform, and its associated problems in regulatory controls, but had not yet chosen or been required to take responsibility for the repercussions of moderation on its employees. This thesis argues the settlement was the most significant social media response to trolling, in the acceptance of responsibility for human suffering facilitated by the service. This paves the way for similar claims by trolled journalists to seek compensation from their employers.
International moderators excluded

The location of moderators determined their inclusion in the compensation payment. The preliminary settlement only covered Facebook content moderators working in California, Arizona, Texas, and Florida from 2015 until 2020. Due to the extensive volume of content, Facebook contracted its moderation work to other companies, with many of the contract companies hiring their workforces from developing countries where labour is cheaper, including India and China. The settlement payment excluded many content moderators who were performing the same tasks and suffering the same mental health issues as those in the four American states and does not cover people hired beyond Facebook to moderate social media content. This illustrates weaknesses in Facebook’s content moderation system that could mean trolled journalists are left exposed.

Resilience

In response to a further recommendation from the settlement, Facebook required contractors to “screen applicants for emotional resiliency as part of the recruiting and hiring process” (Newton, 2019a, para. 14). In this study’s in-depth interviews with trolled journalists (in chapter 4.2 on p. 64), seven of the ten highlighted the need to be “resilient” as a journalist. Additionally, all ten participants labelled trolling a “significant” workplace problem, described the suggestion to ignore trolls as an “inadequate” coping strategy, and suggested that relying on resilience is harmful to the journalist (as reported on p. 72 and discussed on p. 79). Facebook’s continued reliance on emotional resilience as a coping strategy for content moderators as still suggested as questionable. The same applies to newsrooms requiring journalistic resilience in place of effective strategies to reduce trolling.

Legal precedence

In the unprecedented compensation case, plaintiff lawyer Steve Williams (cited in Newton, 2020) said it signified that “the harm that can be suffered from this work is real and severe” (para 3). Although the class-action lawsuit was filed in a Californian state court in September 2018, the case opens up the possibility for
further legal actions to be filed in other countries, including Australia. Legal responses are discussed and critiqued later in this study (in chapter 4.7 on p. 265).

This chapter describes Facebook's responses to trolling in five categories: support strategies, feature strategies, functional strategies, research strategies and corporate approach strategies providing clarity about an important stakeholder in the trolling paradigm. The next chapter will investigate responses by Australian newsrooms. While social media platforms sit at the junction of governance and support levels in the systems model, newsrooms that once sat at the support level because they have limited control over the behaviour of people on social media platforms, are beginning to sit at the junction of governance and support level. Newsrooms have begun to institute governance procedures such as social media and workplace health and safety policies, in addition to training and support, that contribute to the online safety of working journalists.
Chapter 4.6
Newsroom Responses

Results and Discussion

A timeline of Australian newsroom responses

Question E - What strategies were implemented by Australian news organisations in responses to trolling?

Trolling has cemented itself as a fundamental part of the current Australian news media landscape. Interactivity provided by social media is a significant contributor to online news’ popularity. However, this interactive function of Web 2.0 has also facilitated trolls’ access to their victims. Comment sections on reputable news sites routinely burst with insults, presenting a significant workplace problem for Australian journalists (Barnes, 2018; Citron, 2016; Jane, 2015; Phillips, 2015). An online culture has emerged where vicious and frequent trolling has become commonplace, and where rape threats have become a predictable default response to journalistic practice (Bartlett, 2015; Greig, 2016; Hunt, 2016; West, 2015; Ford, 2016). Although trolling is evident on a global scale, the academic literature on this phenomenon is still developing in 2021.

Although comprehensive analysis of international news organisations fell beyond the scope of this study, which specifically focused on the trolling of journalists in the Australian news media, notable global newsroom responses were included in the data set and relevant discussion sections where appropriate for contextual analysis. The study selectively analysed three popular Australian newsrooms of differing ownership: The Sydney Morning Herald, News.com.au and ABC.

The research design strategy that addresses Question F: What strategies were implemented by Australian news organisations in responses to trolling? applies mixed methods applied to archival evidence, thematic critical analysis and
interviews. As a starting point, the chapter used the 528 articles in the thematic content analysis from chapter 4.1 (on p. 39). Twenty-four of these articles mentioned newsroom responses to trolling and were deemed relevant for inclusion. Google was then used as a search tool with the search inquiry: “troll” + “newsroom” + “response” + “Australia” retrieving 8,540,000 results. On closer analysis of the first ten pages totalling 100 articles, 17 additional articles were deemed relevant for inclusion. From thematic analysis of the 41 articles retrieved as archival evidence, newsroom responses that documented changes to news organisations websites, social media pages and comment streams, in addition to support and training programs are listed in chronological order. To provide further primary data, media academic Dr Jennifer Beckett whose research explored the safety of online moderators was also selected for the interviews conducted in 2020. In addition, journalist and academic, who in 2019 completed her PhD *Destroying the joint: A case study of feminist digital activism in Australia and its account of fatal violence against women*, that presented a history and assessment of the online feminist movement, Dr Jenna Price, was selected to interview. Both interviewees are Australian. Respondents were reached through academic and industry contacts. The semi-structured interviews went for approximately one hour and provided both quantitative and qualitative data. Through predominantly open-ended questions, robust qualitative data was acquired. The study identified emerging patterns regarding the use and effectiveness of newsroom support systems, which were drawn upon for comparison, discussion and critique. Newsrooms *ABC, 9 News Perth, Seven West Media*, and *Network Ten* in Subiaco were contacted and asked about their policies in response to trolling. *ABC* policies were found on their website. *Seven West Media* and *Network Ten* dealt with matter internally on a case-by-case basis, and *9 Perth* (formerly *Fairfax Media*) refused twice to provide a copy of their handbook for online safety produced by Julie Posetti. While interviews with newsroom executives may have yielded more first-hand information, it was beyond the scope of this thesis.

**Brief definition of key terms**

In newsroom contexts, trolling is part of the broader issue of management of user
generated content, including screening for inappropriate content. However, the definition of inappropriate content is not universally recognised and can differ between various news organisations. Individual newsrooms commonly have guidelines on their websites and social media pages that users must adhere to when engaging online and that moderators can reference and implement when assessing content. The boundary between appropriate and inappropriate content is often left up to individual interpretation. While racist comments are usually flagged for removal as they are unlawful under section 18C of the Racial Discrimination Act 1975 (Cth) and Racial Hatred Act 1995 (Cth), tolerance of a wider spectrum of content including profanity, threats, and insults vary depending on the individual publication (Chadwick, 2019).

**Brief background**

While newsroom interactivity in the form of letters to the editor have long been a feature of newspapers, there are competing claims about the first news website to offer opportunities for dialogic interactivity between news content creators and their audiences. In 1997, *ABC* launched its first online discussion forum on The Frontier website (*ABC*, 2021). Another early pioneer was Bruce Ableson who in October 1998 introduced reader comments to his blogging platform Open Diary that allowed users to respond to each other. In the same month, software developer Dave Winer created a discussion board and published its first comment on October 5, 1998: “too bad coders can’t be like rock stars and get their money for nothing and their chicks for free” (Erard, 2013, para. 6). Integration on news websites began later the same year, with American publication *The Rocky Mountain News* credited as one of the first newspapers to add the feature (Masullo Chen & Pain, 2017). Australia’s *ABC* was years ahead of most Australian media in its adoption of dialogic interactivity and had a rich history of user commenting from 1997 onwards (*ABC*, 2021). Although the *ABC* opened a limited selection of opinion and analysis stories for comment in their ‘The Drum’ section, Martin (2015) described how the *ABC* were under political and regulatory pressure to maintain news impartiality and had prioritised public sector expenditure on reporting rather than moderation costs. However, Martin (2015) added that the *ABC* sought user news interaction through other strategies including local station
and current affairs websites, and innovative social broadcast hybrids such as their ‘Q&A’ program which incorporated live moderated Twitter feeds. Globally, newsrooms began to embrace the two-way communication model in a gradual progression until a swift spike in uptake that included a renewed focus on it by the ABC in November 2005 (ABC, 2005), followed by The Washington Post in June 2006 and The New York Times in November 2007. Between 2007 and 2008 there was a 42 percent growth in the number of top circulating news sites with comments sections in the United States (Santana, 2011). By 2008, 75 percent of the top 100 most circulated newspapers in the US had comments sections (Santana, 2011) which trolls used to abuse journalists. In a 2012 report for the Press Council, Martin and Dwyer (2012) described the emerging need for comment moderation:

As online contributors are not necessarily aware of the codes of practice and standards that govern, however imperfectly, professional media work, content moderation is an essential part of ensuring content accuracy, authenticity and ensuring civility in comments and discussion spaces.

(p. 22)

Although they suggested the primary motivation of moderation was to “protect corporate reputation” (p. 22), they also suggested its design could ensure user trust in industry standards by removing inaccuracies and constraining ethically problematic behaviour such as trolling.

4.6.1 Qualitative data results of newsroom responses

From thematic analysis of the 41 articles retrieved as archival evidence in this chapter, emerging key themes were grouped to provide ten newsroom responses to trolling in the Australian news media. Changes to the interface and functional features of gated news websites, industry-led campaigns, counselling resources, and training procedures were included. The ten newsroom responses were grouped into four stages, based on their key objectives. The newsrooms’ responses progressed from basic website feature upgrades to workplace procedural changes that considered the emotional and behavioural impacts of trolling. These ten newsroom responses are listed in chronological order below.
Stage One: Comment stream regulation

1. Newsrooms employed comment stream moderators (2005)
2. Newsrooms began to shut down comments in forums (2002) and comment streams under news articles (2012)
3. Newsrooms removed comments on selected articles (circa 2012)

Stage Two: Activism

4. #StopTheTrolls campaign was launched by The Daily Telegraph (2012)

Stage Three: Tightened security protocols through gated websites

5. Newsrooms authenticated users before commentary (2013)
7. Newsrooms may have deliberately redacted the names of journalists from selected articles (circa 2014)
8. Newsrooms introduced technology assisted moderation (2015)

Stage Four: Training and support

9. Newsrooms acknowledged psychological impacts and provided counselling
10. Newsrooms introduced trolling specific training resources

Each of the ten newsroom responses were further investigated using the Google search engine to provide 21 supplementary articles that were not part of the initial data set. This collection of 62 documents was used as a base for the research and was drawn upon to provide the qualitative descriptions of each individual newsroom response.
4.6.2 Thematic critical analysis

The next step was a thematic critical analysis (Weerakkody, 2015). The responses are individually described and then discussed in chronological order. In a collective narrative, each newsroom response is individually critiqued. Through the grounded theory method, additional themes materialise and are discussed within the textual timeline of the progressive commentary. Emerging themes within this timeline are integrated for discussion to provide a robust commentary for contextual analysis for a systematic and vigorous evaluation. The thematic and critical analysis is predominantly qualitative; however, quantitative evaluation of time periods are also incorporated.

Stage One: Comment Stream Regulation

The first three newsroom responses were grouped in Stage One, addressing the need to monitor, regulate, restrict or remove the comments section below published articles on news websites. The study coded these strategies as reactions to the increased frequency of trolling in the Australian news media.

First newsroom response:

Newsrooms employed comment stream moderators (2005)

Identified as the first response, journalists, who were often first-year graduates, were employed by newsrooms to moderate user content within comment streams below published articles on news websites. ABC first policy that mentioned the commentary moderation in its 2005 House Rules.

• 2005, November 2

We reserve the right to remove any comment or user from our blogs, although we have no duty to do so. When comments are deleted, it's usually because they violate one or more of our house rules. Repeated violation of these rules may result in the suspension of your commenting privileges.

(ABC, 2005, para. 1)

Seven years later, Fairfax Media employed Rob Ashton as its first full-time comment moderator in 2012. Inundated with comments and responding to increased public pressure in 2016, Fairfax Media employed additional staff to the equivalent of 3.5 full-time comment moderators to monitor its five metropolitan news websites, including both smh.com.au and theage.com.au in 2016 (Ashton,
Discussion of this newly formed role follows.

The comment stream section of online news providers enabled millions of readers to share and discuss their opinions on news topics of interest. This increased engagement amplified readership. Moderators aimed to facilitate respectful and informative discussions by deleting insults, defamation, and hate speech, in addition to unverifiable facts (ABC, 2005). In some parts of the ABC and Fairfax Media organisations journalists moderated their own comments. Some comment moderation was done before comments were published, and some after. In order to keep the community engaged, the moderation process is required to be transparent and comprehensive; however, news providers must not give the impression of censorship or dissemination of fake news. As manual moderation is expensive, the process becomes increasingly unfeasible with cumulative discourse. It is from this two-way, interactive function of comment streams that trolling emerged. Its impacts on journalists employed as comment stream moderators have been documented in chapter 4.5 (on p. 127). Risch and Krestel (2018) suggested comment moderators were “the only known way to prevent trolling attacks, to watch the compliance of users “netiquette” and to keep up good discussions” (p. 166). While the other solution is removal of the comment streams, this would limit the platform’s capacity to host “good discussions.” To further unpack emerging difficulties with the moderation and governance of social media in response to trolling, an interview was conducted with Dr Jennifer Beckett, a lecturer and media academic, whose research focused on online community management. Excerpts from this primary data are presented in section 4.3.1 (on p. 70).

The heavy reliance on comment moderators created issues over time, and human moderation methods proved a lengthy process. Time delays between a reader posting a comment and it appearing on the news website became apparent. At the conclusion of 2017, The Sydney Morning Herald expressed concerns with the quantity of comments to be moderated preceding a year of active debate surrounding same-sex marriage, Donald Trump’s first year of presidency, and the conviction for rape of international film producer Harvey Weinstein and
accusations of sexual assault against Australian television presenter Don Burke. Ashton (cited in *The Sydney Morning Herald, 2019*) stated, “some of you have understandably complained about delays in moderating your comments. We will continue to do our best to moderate them promptly. But sometimes we may not have time to moderate at all. I cannot guarantee that we will be able to keep up” (para. 5). *The Sydney Morning Herald* revealed that comments unable to be reviewed by moderation staff were discarded (Ashton, 2019).

**Second newsroom response:**

*Newsrooms began to shut down comments in forums (2002) and comment streams under news articles (2012)*

Academic debate ensued concerning decisions to keep or remove comment streams on Australian news platforms, emerging in July 2000 when Margo Kingston founded Webdiary for *The Sydney Morning Herald* (Webdiary, 2020). “What was once 'an open conversation' between Margo and her readers became an open conversation between the readers” (para 2). Prior to this, letters to the editor had been a key part of Australian newspapers that aired public debate on major political and social issues, and were they the main channel for readers to respond to the content they consumed (Smith, 2017). Two key articles by Green (2018) and Syfret (2019) described why more news websites were dumping comment streams and these articles form the basis of this discussion section. From the early 2000s, media outlets began publishing more of their content on online, and many editors and reporters were cautiously optimistic that providing a space for online commentary would elicit more diverse audience engagement and create stronger connections between content creators and consumers (Green, 2018). Two significant factors that differentiate between letters to the editor and online commenting are moderation and selectivity. Providing adequate oversight of user commentary was costly and yet defamation risks meant it was required. Many news outlets found the commentary fell far short of the initial goal of encouraging debate and civil discourse among a representative selection of users (Syfret, 2019).

After experimentation and discussion, many news organisations concluded
comment sections did not provide a useful experience for the vast majority of users and became overwhelmed by anonymous contributors who too often hijacked comment threads with offensive and inappropriate submissions (Green, 2018). Jensen (as cited in Green, 2018), who suggested public engagement is a key part of media organisations and must continue to be to keep the business model afloat, considered the decision to shut down comment streams as a hard-line proactive response to trolling by newsrooms. Critics in further literature including Risch and Krestel (2018) suggested a semi-automatic, holistic approach is the optimal strategy, with news providers including comment features but also providing context, which includes transparent information about commentators.

Of the three Australian news websites analysed in content analysis of this chapter, only the ABC had removed its commenting function. The remaining two had active comment streams at the conclusion of 2020, which were The Sydney Morning Herald and News.com.au. The ABC emerged as the leader in Australia in its hard-line response to trolling, identified as the first newsroom globally to remove comments from forums in 2002 (Martin, 2012) and from comment streams under news articles in 2012. In doing this, it followed the BBC which cut access earlier from news due to budget cuts in 2011 (Plunkett, 2011). Although the scope of this study considers the Australian news media, notable international newsrooms that were mentioned in the 62 articles analysed in this section were included to provide context for comparison and discussion. This chapter identified 22 newsrooms that had removed comment sections from their news websites in response to trolling, of which two were Australian: ABC and Mumbrella, and 16 were American. However, after this chapter’s research period had concluded, Mumbrella reopened its comments section for a six-month trial in November 2021 (Mumbrella, 2021). Multiple responses from each Australian newsroom are listed to show the evolution of responses, in which trial periods were adopted. Each newsroom response is listed in chronological order below.
• 2002 – ABC’s first response
(Australian Broadcasting Corporation)
  o ABC first began to shut down commenting on forums (Martin, 2012).

• 2012 March 16 – ABC second response
(Australian Broadcasting Corporation)
  o ABC directed viewers of private and commercially owned social media platforms to make comments on its shows and news articles on their social media accounts instead of on its website, closing the ABC-moderated message board (Knox, 2012).

• 2013, September 24 – Popular Science
(American quarterly magazine)
  o Popular Science’s Director of online content, Suzann LaBarre, announced the decision to remove the comments section saying, uncivil comments not only polarized readers, but they often changed a participant’s interpretation of the news story itself stating, “we’re shutting off comments” (LaBarre, 2013, para. 5).

• 2013, December – USA Today’s
(American daily newspaper)
  o USA Today’s Director of content development, Jamie Mottram, (2014) stated the decision had not reduced their audience reach: “when comments were removed in December 2013, it had 8 million monthly unique users, (11 months later) in November 2014, it had 17 million monthly unique users, and the time spent per visit had not dropped” (Mottram, 2014, para. 1).

• 2014, April 6 – Vox Media
(American news website)
  o Vox.com launched in 2014 without a comment function, stating that “flame wars turn readers off” (AFP, 2015, para. 6).
• 2014, April 12 – The Chicago Sun-Times
(American daily newspaper)
  o Managing editor Craig Newman announced The Chicago Sun-Times would temporarily shut down comments as “these forums too often turn into a morass of negativity, racism, hate speech and general trollish behaviours that detract from the content” (Graef, 2014, para. 4). He promised a “new commenting system” (para. 1).

• 2014, August – CNN
(American news-based pay television channel)
  o CNN “quietly disable[d] comments on most articles [published on cnn.com] during the protests in Ferguson, Missouri” (Finley, 2015, para. 7). However, some comment sections are selectively activated on stories that editors feel have the potential for high-quality debate, and when writers and editors can actively participate in and moderate those conversations (Gross, 2014).

• 2014, November 7 – Reuters
(International news organization established in London)
  o Reuters’s executive editor Dan Colarusso (2014) announced, “Reuters.com is ending user comments on news stories” (para. 2). Colarusso suggested social media platforms were the successors of the commentary space.

• 2014, December 17 - Mic
(American media company)
  o Mic’s editor Christopher Altchek (2014) announced the removal of comments from mic.com with the explanation, “rather than assigning team members to manage the comments on our site, we are investing our engineering and editorial resources in new products and storytelling formats that benefit our audience” (para. 3).
• 2014, December – Recode
(American technology news website)
  o Recode’s executive editor Kara Swisher announced that comment streams would be removed from recode.net and encouraged a migration of discussions to social media platforms (Swisher & Mossberg, 2014).

• 2014, December 15 – The Week
(British and American weekly news magazine)
  o The Week’s editor Ben Fruman (2014) announced the removal of comments, explaining “too often, the comments sections of news sites are hijacked by a small group of pseudonymous commenters who replace smart, thoughtful dialogue with vitriolic personal insults and rote exchanges of partisan acrimony” (para. 2).

• 2015, January 27 – Bloomberg Businessweek
(American weekly business magazine)
  o Bloomberg.com website was relaunched with no comments. Digital editor Josh Topolsky explained, “I have looked at the analytics on the commenting community versus overall audience. You are really talking about less than one percent of the overall audience that’s engaged in commenting… in the grand scheme of the audience, it does not represent the readership” (cited in O’Donovan, 2015, para. 9).

• 2015, July 6 – The Verge
(American technology news website)
  o Vox Media’s online news website, The Verge said it was “turning off comments for a bit” (Patel, 2015, para. 1) as the tone of dialogue was “getting a little too aggressive and negative” (para. 2).
• 2015, July 27 – *The Daily Dot*

*Amerci-an news website*

- In a joint announcement, *The Daily Dot’s* founder and editor-in-chief Nicholas White and former managing editor Austin Powell announced the removal of comments on *dailydot.com* stating, “the system has been difficult to manage” (Powell & White, 2015 para. 3), and further adding, “we suspect that many publishers will soon find that their existing commenting systems do not serve their readers as the conversation continues to move off websites to social media, where most of our content is discovered and consumed” (para. 3).

• 2015, September 8 – *24.com*

*South African largest English language online news publication*

- Following moderation challenges, 24.com removed its commenting function. Editor-in-chief Andrew Trench (cited in *News24*, 2015) said, “While *News24* is an active proponent of freedom of speech, we believe the comments sections on our news sites are not adding significant value to our content offering and the user experience” (para. 4).

• 2015, October 5 – *Vice Motherboard*

*Canadian and American news website*

- *Vice Media’s* website *motherboard.vice.com* decided to permanently remove commenting, concluding “the scorched earth nature of comments sections just stifles real conversation” (AFP, 2015, para. 3.). It announced that recommenced letters-to-the-editor would be filtered by moderator staff.

• 2015, October 6 – *Reddit*

*American social news aggregation*

- Reddit launched its news site *Upvoted*, which has no commenting feature (Finley, 2015).
• 2016, August 23 – National Public Radio (NPR)
(American privately and publicly funded non-profit media organization)

  o On September 28, 2008, NPR introduced its reader commenting system, an option it embedded through a third-party system at the end of most articles on the site. The then-editor Dick Meyer stated, “NPR was late to this game” (2008, para. 5) and that he hoped the dialogue exchange would be “smart and generous of spirit” (para. 2). However, the toxic culture grew over time and, in 2016, the former managing editor Scott Montgomery stated it no longer “provided a useful experience for the vast majority of users” (para. 1).

  o Eight years after launching, the comment section was removed.

  o The removal did not significantly impact its readership and NPR found only a small proportion of its visitors were commenting on articles (Guaglione, 2017). After being comment free for one year, NPR’s interim managing editor Sara Kehualani Goo stated, “the number of users for the May-to-July period grew 18 percent in 2017 compared to the same period a year earlier” (cited in Guaglione, 2017, para. 11). However, Facebook comments for NPR posts were continuing to steadily increase, reaching a high of 700,000 in January 2017. This may suggest readers who previously shared opinions on NPR's website had migrated their commentary to Facebook, and therefore the newsroom response had not decreased the trolling frequency, it had simply redirected the location of trolling content to Facebook.

• 2017, April 14 – The Daily Beast
(American news publication)

  o The Daily Beast temporarily removed commenting from the thedailybeast.com; however, 7 days later, on April 21, 2017, it launched a new comments platform. The software produced by LiveFyre aimed to ensure a higher level of discourse and less vitriol through the integration of “better moderation tools and a low
system tolerance for spam” (The Daily Beast, 2017).

- 2017, July – MSN News
  (American news aggregator)
  - MSN.com temporarily removed commenting on its sites (Guaglione, 2017).

- 2017, August 30 – Al Jazeera English
  (Qatari state-owned news channel)
  - Al Jazeera English disabled commenting, stating, “the comments section was hijacked by users hiding behind vitriol, bigotry, racism and sectarianism. The possibility of having any form of debate was virtually nonexistent” (Guaglione, 2017, para. 3).

- 2017 September 7 – ABC’s third response
  ABC relaunched commenting in a three-month trial of Livefyre comments on abc.net.au, which aimed to:
  - Offer comments on a wide selection of content, but with a focus on the stories that were generating the most interest on the ABC News site.
  - Build a community of commenters interested in sharing their perspectives on the news of the day in a civil way.
  - Highlight the best of this content and surface it for all ABC News readers, even if they are not commenters.
  - Determine whether comments on our site are still valued, despite the rise of Facebook (Kemble, 2017, para. 17).

- 2017, October 9 – Independent
  (Irish news website)
  - Independent News and Media (INM) removed the comments section from its main news website Independent.ie to “minimise the legal risk” (Hamilton, 2017, para. 1) to its business, citing Ireland’s “draconian libel awards system” (para. 2) as the primary driver
behind the decision.

- 2018, February 15 – ABC’s fourth response

After five months, ABC again ceased the comments section on abc.net.au concluding that “despite technological advances, moderating comments is still (too) labour intensive” (Kemble, 2018, para. 32). ABC acknowledged that comments are a valuable way for readers to share their experiences and still play an important role in the social media world; however, the toxic influx of discourse, particularly regarding politics, distracted from the interactions. ABC disclosed the most provocative articles account for the bulk of the comments and revealed the top five:

1. Same-sex marriage bill: We need amendments and it’s not just about bakers and florists with 3,946 comments.
2. Donald Trump hits out at 'shithole' countries amid talks on protections for immigrants with 3,071 comments.
3. Donald Trump’s obsession with Hillary Clinton is a sign of a deeper problem with 2,734 comments.
4. Steve Bannon’s apology masks the chatter about Donald Trump and the 25th Amendment with 2,667 comments.
5. Sam Dastyari warns white nationalism on the rise after pub ambush by far-right group with 2,487 comments.

(Kemble, 2018, para. 21).

- 2020 January 29 – Mumbrella first response

(Australian marketing and media industry news website)

- In a partial ban, Mumbrella’s head of content Tim Burrowes (2020) said, “under our new moderation policy, we will only be publishing comments that we believe are worth our audience seeing” (para. 3).

- 2020 September 18 – Mumbrella second response

- Seven months after their first announcement, Mumbrella’s head of content Damian Francis (2020) announced the website would “no
longer publish reader comments. This policy will be for all of the content we publish, regardless of what it is” (para. 1).

The *ABC* emerged a leading player in a hard-line response to trolling and was identified as one of the first newsrooms globally to start removing comments from their forums in 2002 (Martin, 2012) and further announce the removal comment streams on March 16, 2012, within the *social problem period* (described in section 4.1.3 on p. 45). However, on September 7, 2017, *ABC* launched a trial of Livefyre comments on abc.net.au, following *The Daily Beast*’s introduction of the technology five months earlier on April 21. Livefyre (now owned by Adobe) is a commenting platform founded in 2009 that incorporates marketing tools to increase engagement of user-generated content for companies including CNN, The Coca-Cola Company and Major League Baseball (Lardinois, 2016). However, after a trial period that lasted five months, *ABC* again ceased all commentary concluding “despite technological advances, moderating comments is still (too) labour intensive” (Kemble, 2018, para. 32). The trial displayed a willingness and openness to adapt to the contemporary two-way communication environment, and demonstrated the *ABC* valued some components of the commentary format enough to attempt its reintroduction. However, the influx of toxic discourse proved overwhelming for the technology, and commentary was removed indefinitely. Both time pressure and financial restraints were cited as compounding tensions that burdened moderators beyond the already recognised psychological impacts. In support of the *ABC*’s claims, an additional six newsrooms attributed a lack of resources to deal with the unmanageable quantity of content as the main contributing factor towards the removal of commentary (*Mic*, 2014; *The Verge*, 2015; *The Daily Dot*, 2015; *24.com*, 2015; *NPR*, 2016; *Mumbrella*, 2020). Newsrooms’ resources were concluded to be better spent producing editorial content rather than monitoring reader dialogue.

Although comment sections can provide a place for readers to share opinions, engage with others, and contribute to topics that affect their communities, these sections can also provide a platform for toxic discourse marked by disrespect, profanity, and yelling in all caps. Houston, Hansen and Nisbett first documented
this evidence in a 2011 study *Influence of user comments on perceptions of media bias and third-person effect in online news*. Thematic analysis of the 41 articles retrieved as archival evidence in this chapter revealed that the five newsrooms that removed the feature asserted that trolling comments on their news website adversely affected the credibility of the news organisation and uncivil comments not only polarised readers, but often changed a participant’s interpretation of the news story itself (Popular Science, 2013; The Chicago Sun-Times, 2014; CNN, 2014; The Daily Dot, 2015; Al Jazeera English, 2017). In support of the identified claims, researchers from the University of Texas (cited in Masullo Chen & Lu, 2019) concluded, “people who viewed news stories with only uncivil comments had less positive attitudes toward the site, and saw it as less valuable compared to those who saw stories with only civil comments” (p. 2).

The most frequent clarification, explanation or excuse for the removal of comment streams was the progression to social media platforms. In conjunction with reported strains on internal staffing, 11 newsrooms publicly passed the responsibility to Facebook and Twitter to regulate conversations away from their own news websites (ABC, 2012; ABC, 2017; Popular Science, 2013; CNN, 2014; Reuters, 2014; Mic, 2014; Recode, 2014; The Week, 2014; The Verge, 2015; The Daily Dot, 2015; NPR, 2016; Al Jazeera English, 2017). This finding revealed 50 percent of all newsrooms examined via the documents gathered for this section that removed comments publicly declared that trolling had escalated beyond their ability to control it and handballed the problem to the social media companies to deal with and bear legal responsibility for or to choose not to deal with, independently. *The Week’s* editor Ben Fruman (2014) stated, “it is no longer a core service of news sites to provide forums for these discussions” (para. 3). His sentiments were echoed throughout all eleven newsrooms listed in this section. In contrast, *ABC’s* John Kemble (2017) stated, “decisions to ditch comments are often blamed on the rise of social media, Facebook in particular, as the go-to place for audience interaction” (para. 6). However, he further criticised the wave of deflection suggesting, “dig below the surface and it is often a case of resourcing. Building a community is difficult and building the sort of community you want is even harder” (para. 7). Kemble’s suggestion that deflecting the responsibility to
social media emerged as a convenient cover and distraction from the true depth of the trolling problem is valid. More accurately and more likely, the underlying problem with trolling in newsrooms emerged from an accumulation of resource concerns, including strained staffing, insufficient finances, and inadequate technical aptitude (Green, 2018). In unison, these deficiencies meant that newsrooms could no longer cope with trolling. Publicly deflecting responsibility to social media organisations did not reflect underlying motives or acknowledge the true scale of the trolling problem. Despite suggestions from the eleven newsrooms that social media was better suited to comment regulation, Facebook and Twitter were not exemplary regulators at the time. Newsrooms simply could not cope with the onslaught of vexing commenters through internal means.

The presence of comment streams on some news websites illustrates the difference of opinions in the current media industry. Debate still ensues about whether the positives of free speech and reader engagement outweigh the cost of relentless vitriol. Some publishers including *The New York Times* have been able to effectively maintain commentary through robust moderation and technology. *Vox Media* journalist who oversaw The Coral Project, Andrew Losowsky (2018) suggested:

> There are compelling reasons why it is worth investing in comments on news sites. While they are usually a small percentage of the total audience, commenters are often the most loyal and most valuable readers. They spend longer on the site, they come back more often, they share more links to the site, and they are more likely to pay for subscriptions.

(Para. 6)

However, positive romanticisms have become far less common, replaced with critical inquiry. Although it resisted shutting down its comments, Australian digital news service *Crikey* began to openly question the sustainability of constant moderation under financial strain by publishing the article *Is it time to kill the comments section?* (Syfret, 2019). Although trends were beginning to suggest that more newsrooms are choosing to opt out of comment sections, active professional and academic debate continued. *The Daily Dot* (cited in Powell, 2015) eloquently concluded in favour of comment streams, “of course we would like to see a more civil, compassionate Web, but we want to be careful that in the name of fostering
civility, we do not inadvertently kill all dissention” (para. 4). An additional argument for comment streams may stem from news companies’ desire to boost metrics about time users spend on their sites as these metrics are used to encourage advertisers to buy space.

**Third newsroom response:**

**Newsrooms removed comments on selected articles (circa 2012)**

In interview with *The Sydney Morning Herald* (Raggatt, 2013) on August 15, 2013, the Press Council chair, Professor Julian Disney, confirmed that some newsrooms had begun selectively removing the commenting section of potentially problematic articles, stating:

- “Cautious publishers were increasingly limiting the number of articles they opened for reader comments because of abusive posts” (para. 2).
- “There had been a notable decline in the number of ‘open’ articles compared to two years ago” (para. 3).
- "The quality of comments that are being received on a lot of streams have led newspapers to just not open them on many articles” (para. 4).

These comments were the first identified data that documented Australian responses to trolling. Disney suggested this decline had occurred over two years. Therefore, the study has allocated this response to have emerged between 2012 and 2013.

As comment moderation became a round-the-clock and resource intensive process, even large teams found it difficult to police at scale. Some newsrooms attempted to navigate the problem by only opening commentary for a limited number of articles. In 2016, *The New York Times* removed reader comments from 90 percent of articles that the publication considered a higher risk for inviting vitriolic dialogue (Etim, 2016, September). In supporting evidence, on February 1, 2016, *The Guardian* announced the same strategy. Executive editor Mary Hamilton (2016) stated:

- “*The Guardian* took the decision to cut down the number of articles with open comments on stories relating to a few contentious subjects, particularly migration and race” (para. 2).
• “Improving comments does not mean censorship or eliminating criticism” (para. 1).
• “The aim is not to stop comment appearing, but rather to enable us to management them more effectively... and make sure the discussion is constructive and not abusive” (para. 2).

*The Guardian* stressed the importance of being responsible hosts who govern constructive debate. The prevailing tone of proceeding commentary is set by the article topic; therefore, proactive measures can be identified and implemented before trolls target flagged stories.

In Raggatt’s (2013) article that provided the first documented data of comment removal on contentious articles, opposition was also noted. Rob Ashton (2013), a comment moderator at *Fairfax Media*’s five Australian news websites, was cited as saying, “there was no move to open fewer articles to reader comments” (as cited in Raggatt, para. 12), and further suggested although “articles on controversial topics such as Muslim immigration, same–sex marriage and asylum seekers may have a higher rejection rate, that does not mean we should not open them” (para. 15). At the time, *Fairfax Media* advocated a strong stance against the removal of commentary, reflected in continued commenting across their ownership web, including the *Canberra Times*, whose chief online producer Lucy Rickard actively invited reader engagement and encouraged comments to generate conversation (Raggatt, 2013).

**Stage Two: Activism**

The fifth newsroom response was coded as the first attempt to proactively address the rising frequency of online abuse journalists were subjected to as part of their work.

**Fourth newsroom response:**

*#StopTheTrolls campaign is launched* (2012)

Murdoch-owned, Sydney tabloid *The Daily Telegraph* launched the *#StopTheTrolls* anti-trolling campaign on its front page on September 11, 2012. The campaigned pressured Australian politicians and law enforcement to increase regulation and
policing of Twitter in order to identity and prosecute perpetrators of online abuse.
The newspaper front-page feature is shown in Figure 30 below.

**Figure 30**

*The Daily Telegraph take a stand against trolls.*

In a multi-platform campaign, readers were encouraged to digitally sign two petitions and share their support by posting ‘#StopTheTrolls’ on Twitter. The hashtag began to trend across Australia and in 24 hours had been retweeted over 400 times (Lince, 2012). Although the ‘Twitition’ petition is no longer accessible, the Change.org version of the petition received 1,000 signatures in a few weeks. Although some journalists criticised the gradual uptake of signatures as a failure (Posetti, 2012; Ross, 2012), the extensive public awareness generated from the campaign is indisputably a measure of partial success. The swift inclusion of trolling within active public debates spotlighted the gravity of an underlying trolling epidemic that had simmered without generating widespread public concern until this point. The original text of the 138-word campaign published on the change.org website stated:
In a publicly storm, *The Daily Telegraph* published five related articles on their news website over four days: three articles on September 11 (Jones & Byrnes, 2012, *The Daily Telegraph, 2012; The Daily Telegraph, 2012a*), one on September 13 (Jones, Lulham & Frost, 2012) and one on September 15 (Hildebrand & Matheson, 2012). Textual analysis of the five articles yielded examples of highly emotive language that differed from the newspaper’s traditional reporting style, which customarily focussed on facts rather than editorial opinions. Examples of these expressions included “anonymously taunt”, “faceless bullies”, “unmask them” (Jones, Lulham & Frost, 2012, para. 1), “these cowards”, “vile and abusive”, “mercilessly attacking”, “hiding behind keyboards” (*The Daily Telegraph, 2012, para. 1), “helpless victims” (para. 6), “lawless free-for-all” (*The Daily Telegraph, 2012a, para. 3), “anonymous cowards”, “misery and distress” (para. 5), “mindlessly torment”, “grieving, vulnerable or suicidal” (para. 7), “vile and unanswerable” (para. 9) and “condemned the cowards” (Jones, Lulham & Frost, 2012, para. 2). Minimal emotive wording was identified in the fifth article, with formal descriptive language readopted. The initial tonal shift provoked reader interest and sparked discussions about an underlying social problem many had experienced or witnessed in isolation, behind a computer screen. Once private struggles were broadcast across NSW news media and other News Corp publications, they became part of public discourse and awareness.

A deliberate publicity tactic that leveraged celebrity endorsements to fuel the movement is revealed in the thematic analysis of these five articles. The marketing strategy received extensive support from nineteen prominent public figures, who
openly commented on and promoted the campaign. Although celebrity endorsements often require payment, nineteen public figures appeared in the five Daily Telegraph articles analysed. Seven were athletes, two were television presenters, and ten were politicians. The list included NRL players Ben Barba, Josh Dugan, Israel Folau and Robbie Farah, Australian cricket captain Michael Clarke, Ironwoman Candice Falzon, surfer Laura Enever, media personalities Charlotte Dawson and Nathan Jolliffe, then-Prime Minister Julia Gillard, then-acting Prime Minister Wayne Swan, former Prime Minister Kevin Rudd, Opposition Leader Tony Abbott, opposition communications spokesperson and future Prime Minister Malcolm Turnbull, NSW Premier Barry O’Farrell, Attorney-General Nicola Roxon, Communications Minister Stephen Conroy, School Education Minister Peter Garrett, and NSW Police Minister Mike Gallacher. In particular, criticism of political involvement, Ross (2012) doubted the true motives of Stephen Conroy and Nicola Roxon, who were gathering support for a data retention inquiry and had a history of lobbying for legislation that restricted freedom of expression.

Although the significance of celebrity theory has been discussed in depth (in chapter 4.1 on p. 30), further criticisms of these dynamics are explored. Voicing one of these critiques two days after the campaign began, political editor, Simon Copland, published the article, Do we only care about trolling when celebrities are victims? (2012), and suggested that trolling targeted all members of society. The media, reacting at scale only when those in the public eye were affected, diminished the importance of targeted minority groups by comparison. Copland stated, “we have not seen any similar campaigns to stop homophobia, stop racism, to stamp out sexism” (para. 7), which represented the demographics most subjected to prolific abuse both online and offline (Wagner, 2022). However, by virtue of social status, people in the public eye generate interest, and this piqued interested can be leveraged more easily to spark greater interest. The philosophical concept of the episteme used by Foucault (1971) in The Order of Things can be used to contextualise and explain the explosive nature of the campaign. Foucault wrote: “in any given culture and at any given moment, there is always only one episteme that defines the conditions of possibility of all
knowledge, whether expressed in a theory or silently invested in a practice” (p. 183). Therefore, if an issue is raised at a time when the configuration of the episteme prevents its processing it will flounder. But if presented again later when the episteme has lifted it can gain traction. The root problem was not the fact a celebrity or the Prime Minister was trolled; however, the incident emerged from a ‘perfect storm’ that triggered an underlying widespread social issue into the public spotlight.

Australian news media were unanimous on their stance against trolling although opinions about the campaign’s true impact varied between newsrooms of differing ownerships. Although some critical commentary echoed suggestions that media coverage was exaggerated and the campaign was a failure (Posetti, 2012; Ross, 2012), continued media coverage from multiple news outlets including the ABC (Mark, 2012), News.com.au (AAP, 2012) and The Sydney Morning Herald (Yeates, 2012) applied unrelenting pressure on the platform to respond, prompting discussions between the Australian government and the Twitter management team. Nine days after the campaign was launched, changes to the relationship were announced. “Twitter has agreed to hand to Australian authorities the account names and IP addresses of users suspected of bullying or other harmful tweets” (Hutchinson, 2012, para. 1). Twitter’s previous policy had required a U.S. court order before an abusive user account could be identified, an impenetrable obstacle for Australian authorities. However, the agreement did not address or provide any solutions for accounts that are intentionally set up anonymously. These users cannot be identified by Twitter and remain a significant contributing factor to the presence and malice of trolling online.

As discussed in the literature review, raising awareness is a key strategy in addressing wicked problems such as trolling, which are complex, dynamic, and difficult to navigate. Australian Public Service Commission (2007) defined wicked problems as problems that go beyond the capacity of any one organisation to understand and respond to and working across agency boundaries is increasingly important in tackling them. They added that changing the behaviour of large groups of people requires a collaborative approach by multiple stakeholder
groups and poses challenges for traditional approaches, given the fluctuating dynamics of the online environment. The range of traditional levers used to influence behaviour, including legislation, fines, taxes, other sanctions are often part of the solution, but these alone may not be sufficient. More innovative, personalised approaches are likely to be necessary to motivate individuals to actively cooperate in achieving sustained behavioural change.

Thematic analysis of the #StopTheTrolls that ran concurrently with Facebook’s Australian Be Bold Stop Bullying campaign in 2012, revealed bystanders to be an emerging stakeholder group in the trolling paradigm. The widespread public sympathy engendered by the awareness campaign revealed a portion of the population including students, parents and educators that may have witnessed trolling abuse but until this point had taken no personal responsive action (Marszalek, 2012). The campaign galvanised this passive group to respond in collaboration with other stakeholder groups and create a cohesive public stance against trolling. Further scrutiny of trolling came from the non-for-profit sector. The chief executive of mental health advocacy group Headspace, Chris Tanti said, “the campaign is about the bystander, and for the bystander to stop being a bystander and get involved” (as cited in Marszalek, 2012, para. 15). These onlookers or observers emerged as the largest stakeholder group, and by motivating this portion of the public to take interest at scale, sporadic discussions lead to extensive discourse, and widespread media traction was gained.

In summary, the campaign received widespread public praise, international media traction and global recognition, evoking action from a previously passive stakeholder group: bystanders. The campaign was the catalyst for a detectable shift in public perception from anecdotal evidence to a recognised social epidemic that called for urgent and extensive responses from multiple stakeholder groups.

**Stage Three: Tightened security protocols through gated websites**

The next four newsroom responses were grouped in Stage Three, with each response introducing a feature that provided a barrier to comment sections, creating safer gated websites. The study coded the strategies as either reactive
and proactive. Newsrooms globally began to recognise that trolling abuse was not exclusive to the online domain, and sometimes represented a real-life threat to the physical safety of their employees (Barton & Storm, 2014). From thematic analysis of the 62 articles retrieved as archival evidence in this chapter, the 2017 article: *Fighting back against prolific online harassment: Maria Ressa* by Julie Posetti, an academic and former Fairfax Media employee who composed a handbook of recommended newsroom responses to trolling (Ireton & Posetti, 2018) was analysed during this study. The article was focussed on a journalist of 30 years, former CNN war correspondent and the founding CEO and Executive Editor of the social media powered news organisation Rappler, Maria Ressa. Posetti (2017) quotes Ressa calling for tighter security protocols. Posetti describes Ressa as an industry leader who directed a synthesised response to protect female employees who were disproportionately targeted in her newsroom. Ressa was one of the first Executive Editors to openly acknowledge that the harm caused by trolling “must be taken seriously” (para. 17), and had the potential to materialise as an imminent danger in the real world. Ressa asserted that the strategies of blocking, muting, reporting, deleting, and ignoring were no longer adequate, and declared newsrooms must take responsibility for the health and wellbeing of their employees. In support of her claims, two years later, research by Gender Equity Victoria and the Media Entertainment and Arts Alliance (2019) stated “online abuse can move offline when technology is used to access private information that is then mobilised to physically stalk or harm women” (p. 1). The following three newsrooms’ responses were implemented to tighten security protocols by introducing gated features to news websites.

**Fifth newsroom response:**

**Newsrooms authenticated readers before commentary (2013)**

In an additional response, readers were required to authenticate their accounts before any commentary could be published online with *The Sydney Morning Herald* leading by example in 2013. In an attempt to combat the influx of daily abuse, reinstate the gatekeeper function, and reduce the risk of defamation, newsrooms progressively began to implement this strategy, providing publishers with greater control, as displayed in the chronological list below.
• 2013, February 24 – *The Sydney Morning Herald*  
(Australian newspaper)

  o *The Sydney Morning Herald* launched its ‘Comment Page’ on which community editor, Kathyrn Wicks, announced users would be required to log in, using an authenticated account, before contributing to discussions (Wicks, 2013).

Some newsrooms progressively applied the authentication feature both in Australia, and globally.

• 2013, August 22 – *Huffington Post*  
(American news aggregator and blog)

  o *Huffington Post*’s managing editor, Jimmy Soni, announced users would no longer be able to create anonymous accounts, with users’ identification required internally. In further explanation Soni (2013) stated:

    Huffington Post recognises that many people are not in a professional or personal situation where attaching their name to a comment is feasible. This change would not require users to identify themselves in connection with each comment. Rather users are asked to verify their identity when creating an account to reduce the number of drive-by\(^1\) or automated trolls\(^2\).

    (para. 1)

However, the decision to end anonymous comments provoked mixed commentary, reportedly "spark[ing] anger" amongst some users (AFP, 2015, para. 19) yet receiving praise from others. Media expert and sociology professor at Tufts University, Sarah Sobieraj, welcomed the decision in preference to banning all comments stating:

  the ability to speak anonymously online is not just for trolls. Many people turn to the anonymity of the Internet for meaningful dialogue about sensitive topics, such as politics or impropriety in the workplace, that prove difficult to discuss face to face.

---

\(^1\) A user who posts abusive content and immediately leaves the conversation in the digital equivalent of a drive-by shooting.

\(^2\) An automated social media account (also known as a ‘bot’) run by an algorithm, rather than a real person, designed to make posts without human intervention (Barojan, 2018).
American privately and publicly funded non-profit media organisation National Public Radio (NPR) also implemented the strategy, and a visual example of the authentication feature is shown in Figure 31 below.

**Figure 31**

*Unauthenticated readers are blocked from commenting by newsrooms.*

![Unauthenticated readers are blocked from commenting by newsrooms.](image)

(Hugo Rojo/NPR)

An extension of the authentication feature was also identified and described below.

- 2018, October 16 – *The Australian* (Australian newspaper)
  - *The Australian* digital editor, Daniel Sankey, (2018) announced a series of upgrades to the news websites’ commenting functions, stating “when making comments, users will now see a new My Profile tab... users are able to see their comment history, whether a comment is pending moderation, or once moderated, whether it’s been approved or rejected” (para. 2) providing a clear and transparent commentary process.
Accountability has emerged as a key factor to reduce the frequency of trolling, particularly by spam accounts. University of Kent researcher Ian Rowe (2013) attributed authentication to a reduction in the frequency of unfavourable behaviours online. Referencing the Facebook user verification system, the 2013 study found that by introducing accountability, users were “less likely to engage in uncivil discussion” (cited in AFP, 2015, para. 21). This theoretical deterrent did not always succeed in practice. Many newsrooms have not independently adopted this strategy but have linked up with Facebook whose plug-in verifies the identity of users by asking people to use their real names, meaning users are authenticated via the third-party (Yu, 2020). Studies such as the one previously mentioned indicate the Facebook platform authentication process could help set the tone of comments by removing anonymity and providing sufficient information to hold users accountable for unlawful dialogue. However, the authentication process is not fool-proof, and fake accounts are still able to be created and exploited by trolls, with minimal technical aptitude, as was illustrated by Facebook removing 3.3 billion fake accounts in 2018, 6.5 billion in 2019 and 5.8 billion in 2020 (Statista, 2021a).

**Sixth newsroom response:**

**Newsrooms introduced ‘hide comments’ feature for journalists.**

In an attempt to provide journalists with greater control over abusive commentary, some publications followed Facebook’s lead and implemented the hide-comment feature on their websites (Medium, 2015). This optional function allowed journalists to select their preference to conceal all commentary on published articles. This did not remove the comments from being viewed by others but shielded journalists from notifications in a more personalised approach. On November 6, 2015, American blogging platform *Medium* allowed users to adopt this function, announcing:

> We value discussion, discourse, and conversation. We encourage users with similar and divergent viewpoints to engage over the ideas and stories that matter to them. But

---

3 A secondary account for sign-up pages when people do not want to give their primary email address that is often used for posting frequent, miscellaneous, random and excessive content (Sommer, 2018).
we live in the real world and recognise that sometimes you may not want to get in a discussion. Sometimes, you just need to come and post. We welcome that use of Medium as well and want to help you feel safe and comfortable doing so.

(Medium, 2015, para. 1)

Although the retention of comment streams continued to enable trolling, it was reported as a step in the right direction by newsrooms to protect journalists yet still permit open discourse (Reader, 2015). However, the extreme examples of trolling that emerged at the same time (West, 2015), suggest the response was inadequate. Shielding journalists from the growing trolling problem whilst allowing the abuse in comment streams did not contribute to solution. The absence of hard-line action may have contributed to part of the problem.

**Seventh newsroom response:**

**Newsrooms may have deliberately redacted the names of journalists from selected articles (prior to 2017)**

The thematic content analysis in chapter 4.2 identified that, of the 528 news articles, 118 were by journalists whose identities were not included in published articles on smh.com.au, theage.com.au and abc.net.au which may have been in response to trolling, meaning that their names did not appear on their stories. Through quantitative analysis, including Mann-Whitney nonparametric tests, binary logistic regression, and Pearson’s correlation coefficient (in chapter 4.1 on p. 30), this study found the 22 percent articles without journalist by-lines received rape threats, in addition to death threats, sexist comments, feminist comments, body shaming or homophobic comments. The inference we can draw from this by looking at the types of trolling received by male and female victims is that a significant portion of the 118 victims with redacted genders were likely to be female. While there may be other reasons behind not publishing by-lines, if female journalists have fewer published by-lines as a trolling-related strategy, then trolling can be seen to be influencing their style of professional engagement. In acknowledgement of this gender skew, Australian trolled journalist, Ginger Gorman, voiced concerns about the professional and economic harms of trolling to women, stating “gendered cyberhate silences women and constrains their ability to find jobs, market themselves, network, socialise, engage politically and
partake freely in self-expression and self-representation” (Gorman, 2019, p. 71). The redaction strategy did not directly address the gender imbalance that would have been evident when journalists’ names were omitted from the news articles.

Although it is unknown when the redaction process began, it was identified during the data gathering process of the content analysis concluding in December 2017, and therefore began prior to this date. Although no formal announcement has been made public about the strategy, the move would have protected staff from abuse by reducing access to journalists by trolls.

In further evidence supporting the strategic nature of by-line omission, on July 30, 2020, Australian popular culture news website Junkee announced “[we] will no longer publish journalists’ by-lines on some stories to protect them from harassment campaigns (Rigby, 2020, para. 1). This was the first documented evidence of the strategy, although it had secretly been adopted by newsrooms at least three years prior when recognised by this study, if not longer.

**Eighth newsroom response:**

**Newsrooms introduced technology-assisted moderation 2015**

As the trolling problem intensified, large news organisations employed large teams of moderators to weed out inappropriate dialogue, but the quantity of comments often required additional help from external contractors. However, this outsourcing tactic was not viable for budget-stretched newsrooms that could not afford to integrate private vendor software available at the time. They were left to battle trolls internally, further burdening staff who were already time-poor and emotionally strained. Competing newsrooms soon realised they faced a common enemy and combined resources in a collective response. In 2015, *The Washington Post* and *The New York Times* joined forces with technology developer Mozilla to form ‘The Coral Project,’ funded by a US $3.89 million grant from the philanthropic organisation Knight Foundation (Tjaardstra, 2016). The project aimed to create open-source software adapted for news websites to better handle online discussions. After two years, the ‘Talk’ software was launched. The multifaceted approach allowed information to be shared between newsrooms, which fast-
Chapter 4.6 Newsroom responses results and discussion

tracked the technological advancements required to combat the dynamics of trolling online. The Coral Project was bought in early 2019 by Vox Media (Rieder & Skop, 2021) and, according to the company, “powered the communities of nearly 50 publishers in 12 countries” (p. 10), improved commenting practices, and made the use of machine learning more transparent.

By 2016, The New York Times had a moderation team of 10 who collectively reviewed over 11,000 comments per day (Etim, 2016, September) although only 10 percent of selected articles were open for commentary at the time. Reaching critical mass, The New York Times implemented technology-assisted moderation software called ‘Perspective’ in partnership with Google Jigsaw, which had not been “economically feasible until now” (para. 3). A description of the tool was provided by product manager CJ Adams (2018):

> Perspective uses machine learning to make it easier to host good conversations online. Perspective finds patterns in data to spot abusive language or online harassment, and it scores comments based on the perceived impact they might have on a conversation. Publishers can use that information to give real-time feedback to commenters and help human moderators sort comments more quickly. That means news organizations and content publishers can grow their comment sections instead of turning them off, and can provide their readers with a better, more engaging community forum.

(Para. 4)

Using this software to host healthier conversations, The New York Times was able to triple the number of articles with open commentary from 10 percent to 30 percent (para. 5). Newsrooms researched and adopted new technologies to filter out trolling and highlight constructive conversations from readers.

As commentary remained open on most stories, by 2015, The Washington Post received over 22,000 comments per day and struggled to retain a positive tone (AFP, 2015). In response to the overwhelming quantity of comments, the publication also implemented technology-assisted moderation software described in a May 5, 2017, announcement:

> The Washington Post has launched ModBot, a software application that utilizes artificial intelligence to moderate
comments. The proprietary technology uses machine learning to automatically filter comments that require human moderating, flag stories that require real-time monitoring, and approve or delete comments based on The Post’s discussion policy. The technology evaluates comments using an algorithm that has been trained by The Post’s years-long history of human-moderated comments. (The Washington Post, 2017, para. 1)

Working in unison with human moderators, this software recognised stories with high rates of deleted content and flagged the articles for further consideration and potential commentary removal by editors. The Washington Post was the first news organisation to integrate The Coral Project’s Talk software with its own Modbot technology on September 6, 2017.

Australian newsrooms took longer to integrate artificial intelligence software into their moderation models.

- The study noted ABC’s removal of comments in 2012 eliminated the need for any moderation, human or technological.

- Although reporting on the technology used by American newsrooms in 2016 (Sullivan, 2016), Fairfax Media did not integrate any methods and continued to rely on the equivalent of 3.5 full-time content moderators led by Rob Ashton (2016).

- It was not until four years after the American Coral Project commenced, that News.com.au introduced technology-assisted moderation on February 7, 2019. In an unconventional announcement, possibly intended to be humorous, the software was described as: “Trish is the name of the fancy new commenting system here at the website... Trish is not a person – though we do wish she was a semi-retired chain-smoking lady with a blue hair rinse who tolerates no horseplay” (News Corp Australia, 2019, para. 2). Furthermore, the publication did not discuss the significance of trolling-related impacts, with the text of the article insulting another journalist saying, “Trish is a moderator – the queen governing the comments section, like Tony Jones on Q&A but with better hair” (para. 4) and was a poor illustration that underplayed the significance of the social problem.
Although on January 15, 2018, both Microsoft and Ali Baba claimed that their software could read like a human, the potential to deduce content meaning was critiqued as a work-in-progress at the time (as described in section 4.5 on p. 126). But in proof of success in practical application, *Microsoft News* cut dozens of editorial workers on May 29, 2020, as it shifted to an artificial intelligence driven system of picking news stories. Critics questioned the algorithm’s ability to completely replace human judgement (Moses, 2020); however, its implementation proved cost-effective, and the same technology could be used by the news organisation for user-generated content moderation in the near future.

**Stage Four: Training and support**

The final two newsroom responses involved acknowledgement by employers that trolling was a serious issue that required specific workplace training and support. They were coded as both reactive and proactive responses to trolling.

**Ninth newsroom response:**

**Newsrooms acknowledged psychological impacts and provided counselling (circa 2018)**

The psychological impacts of workplace trolling were discussed by Ginger Gorman, Van Badham, Liza-Jane Loch, and Jenna Price in the 2018 Senate Inquiry on the adequacy of existing cyber-bullying laws. Although the inquiry acknowledged online abuse resulted in “serious harms” for journalists (Parliament of Australia, 2018, p. vii), counselling support varied between newsrooms, and there was not yet a universal approach. Findings from research by Gender Equity Victoria and the Media Entertainment and Arts Alliance (2019) stated “a lack of support from employers in these situations (of trolling incidences) led to poor mental health, lack of desire to engage, and low self-esteem” (p. 8), and further claimed “there is no real help in coping with the emotional and professional fall out” (p. 8). The findings reflected Posetti’s (2017) criticisms that “[trolling] damage includes well-documented impacts on emotional and psychological well-being” (para. 26). However, inadequate counselling within newsrooms did not address freelance media workers who are exposed to additional risk. Freelance journalists do not always have access to
institutional resources or support afforded to employees working within structured news organisations. Although in 2015, the Dart Centre for Journalism and Trauma reported that “a coalition of major news companies and journalism organisations endorsed worldwide freelance protection standards” (para. 1), the safety principles did not address online harassment, and left some freelance journalists with limited avenues to seek professional advice. Of particular concern, recently graduated younger journalists were sometimes unable to prioritise the training and equipment that may help keep them safe in their personal budget.

To further examine the adequacy of the counselling and psychological support offered to journalists trolled in the workplace, an interview was conducted with Jenna Price that also addressed the tenth newsroom’s response in the section that follows. These results are displayed in section 4.6.2 (on p. 207).

Tenth newsroom response:

**Newsrooms introduced trolling specific training resources (circa 2018)**

The research that prompted newsrooms to implement trolling-specific training resources arose in 2016. The Media, Entertainment and Arts Alliance published research findings from a survey of over 1000 journalists, most of them women (Media Entertainment Arts Alliance, 2016). The *Mates Over Merit* report highlighted that a significant proportion of female journalists had experienced online harassment, trolling, and stalking during the course of their work, but only 16 per cent of those surveyed were aware of their workplace having existing policies to address online abuse, with another 32 per cent stating that their employer had no such policies (Media Entertainment Arts Alliance, 2016).

While Julie Posetti was unavailable for interview for the duration of this study, a booklet funded by United Nations Educations Scientific and Cultural Organisation’s (UNESCO) and written by Ireton and Posetti (2018) called, *Journalism, ‘fake news’ and disinformation: A handbook for journalism education and training* was obtained for analysis. The booklet consisted of seven modules designed to provide journalists, their employers, and journalism educators framework and guidance to navigate emerging issues in contemporary journalism
practice. The seventh module, *Combatting online abuse: When journalists and their sources are targeted*, is relevant to trolling and was written after Posetti wrote advisor guidelines for *Fairfax Media* on best-practice for newsrooms when trolling occurs in the workplace. The module includes literature that introduces disinformation, defines astroturfing, outlines 13 case examples, lists 12 digital security threats, and provides five basic learning objectives. Strategies were about how to look after their own mental and physical safety. There were no suggestions about lobbying other levels of the system framework to bring about more sustainable change. The handbook has been translated into 23 languages; however, it is not known how much uptake there has been of this module in Australian newsrooms. It is unknown if the *Fairfax Media* training guide may have been a more useful tool as refused to make it available for analysis.

In another UNESCO research project, disturbing recommendations for female journalists were identified. A 2014 report, *Violence and harassment against women in the news media: A global picture*, by Barton and Storm, was obtained for analysis. The report was commissioned by the American-based International Women’s Media Foundation and the English-based International News Safety Institute, part of a broader program of work initiated by UNESCO. The report included concerning advice in response to the harassment of women online. Within the six-page recommendation list, the research suggested:

> Journalists should consider carrying a whistle or a rape alarm in their hand or in an accessible part of their bag, but if flying, they should check beforehand if it can be carried on the airline. Journalists should be aware that some everyday items, such as hairspray and spray deodorant, may be used as weapons and can deter an attacker.  

* (Barton & Storm, 2014, p. 35)

The suggestion by an internationally recognised organisation that journalists carry a rape whistle is of great concern if it is made in isolation and implies that victims are solely responsible for ensuring they are not attacked. However, in this case it is part of a systemic approach to deal with violence against journalists. UNESCO is international and the study recognised that threats of violence against journalists in other countries can be a greater concern, with death and disappearance reports more common in countries such as Russia (Roth, 2018)
Nigeria (Cobus, 2020) and Mexico, which was named in 2020 as the “world’s deadliest country for journalists” (Lakhani, 2020).

4.6.2 Interview qualitative data results

Interview with Jennifer Beckett

In September 2020, an interview was conducted with Dr Jennifer Beckett, a lecturer and academic in media and communications, whose research focused on online community management. Emerging from the second and third Australian newsrooms responses identified by this study, the interview provided key insights into the problems faced by social media moderators. Responses were grouped into key themes that are described and summarised below.

Questioned about the prevalence of trolling in the current Australian news media landscape, Beckett acknowledged the complex problem has become an inescapable and embedded part of modern newsrooms, in which moderators play an essential gatekeeping role:

- Although I would like to think we have reached saturation point, I do not think we have yet.
- Trolling predominantly affects women and people of colour, and as more of these targeted groups end up in online spaces, it becomes more visible.
- If a journalist indicates they are a feminist writer in their by-line, they can expect to be trolled.

Confirming the difficulty of policing comments at scale, Beckett labelled the inability to identify abusers as a serious concern for moderators:

- It is hard to figure out who these people are as they often use a Virtual Private Network (VPN) and scramble their location, making it impossible to track them down.
- Even if a trolling account is deactivated, they can just start a new account with a brand-new name.

---

4 An arrangement whereby a secure, apparently private network is achieved using encryption over a public network, typically the internet (Oxford English Dictionary, 2021).
• There is not exactly a barrier to entry and when that happens, it would make things a lot easier.

Reflecting on the use of comment streams as a streamlined and unfiltered replacement of the former ‘letter to the editor’ process, Beckett suggested political divides further fuel abusive commentary:

• We have become more polarized as a society, and social media provides an unmonitored outlet to say inflammatory things people would not have said in the past.
• The Internet has really thrown out the ends of the spectrum of both left and the right. It is the sexy stuff that gets coverage; the middle does not as much. We are in danger of falling under the thrall of the trolls on both ends of the spectrum.
• It seems to me it is less of a hard-news journalist issue and much more of an opinion or columnist journalist issue.

When questioned about the notion of resilience as a coping strategy for moderators, Beckett described this strategy as an out-dated band-aid solution:

• Resilience is only going to take you so far, and after a while you run out, it is not a bottomless well.
• Employers are basically failing in their duty of care by saying you ought to be more resilient and toughen up.

The commonly advocated technique of ‘don’t feed the trolls’ was discussed openly without being directly questioned:

• The old “don’t feed the trolls” does not mean that they go away, it just gives them less oxygen and they eventually burn out.

Spotlighting content moderators as a crucial yet often overlooked stakeholder in the trolling paradigm, Beckett explained:

• It is also really important to understand that there are a lot of people who work behind the scenes who are not the talent, who are not as visible, and whose job it is to mediate all of this commentary.
Beckett’s own research is being conducted in this space and she suggests further research is required to safeguard social media moderation.

Providing further insights, Beckett explained:

- Trolling is concerning due to the presence of poor general behaviours, and its addictive nature as well.
- Focus needs to be directed towards changing people’s behaviour, and, to achieve that, we need a little bit of cognitive behavioural therapy.
- Popping up cues that make users second guess their actions is going to pick up the majority of people in the middle.
- We focus too much on their behaviour and not enough on modelling good behaviour. That is effectively what community guidelines are there for.

Raising concerns about the heavy reliance on artificial intelligence to resolve content moderation problems Beckett called for more clarity about the difference between artificial intelligence and machine learning. Artificial intelligence refers to programs that can simulate human thought, whereas machine learning is a subset of artificial intelligence that allows applications to learn from data without being programmed explicitly. Beckett said:

- Artificial intelligence is too far off at the moment.
- Current online moderation is not artificial intelligence, it is machine learning, and machine learning lacks the ability to consider context.

When questioned about the emerging bystander theme, Beckett suggested support between users within online communities was a key strategy to reinforce communal standards and weed out trolls:

- If we want people to be good bystanders, we have to tell them how to be good bystanders.
- Part of moderation is called ‘leaving a trace,’ where users are told why a comment has been removed, by referring to the community guideline it has breached. Both the explanation and reiteration of community values forms a sense of ownership over the space.
In summary and discussion, Beckett described the crucial role content moderators play in the function of newsrooms. The frequency of commentary for assessment grew exponentially as global trends opted for online news sources in preference to traditional mediums. Without sufficient alternative means to reduce the quantity of content, the psychological effects on content moderators were outlined, and the requirement for newsrooms to introduce technological assistance to help cope with the workload was inevitable. However, a complete removal of humans from the moderation process was not considered a viable option, and a blended approach involving algorithms and humans was offered as a more practical means of moderation in the current media landscape. Etim (2016, March) advocated the positives of dialogic interactions that introduce readers to other like-minded readers, who may continue to recommend other articles from the publisher to each other:

The news industry has done a reasonable job of displaying readers’ voices in recent years, at least at the bottom of some articles. But not enough thought has been given to elevating them, creating an understanding among readers that their responses can be part of the news.

(para. 4)

Etim (2016, March) explored different ways to reward and engage with good commenters to “create a vibrant online society and protect it” (para. 21). As dialogic journalism is framed as an integral part of the future of journalism practice (Martin & Murrell, 2021), moderators are an important stakeholder in The Trolling Paradigm.

**Interview with Dr Jenna Price**

In September 2020 an interview was conducted with trolled journalist, columnist with the *Canberra Times* and journalism academic Dr Jenna Price. Her work included a collaboration project with Julie Posetti and Fiona Martin about trolling. Emerging from the ninth and tenth newsroom responses of this section, the interview further evaluated the counselling and psychological support offered to trolled journalists in the workplace. Questions were also raised regarding the Senate Inquiry on the adequacy of existing cyber-bullying laws, to which she made submissions that stressed the need for adequate law reform. This aspect is addressed in the next chapter (in section 4.7 on p. 249). Responses were grouped
into key themes that are described and summarised below.

When questioned about the prevalence of trolling in the current Australian news media landscape, Price acknowledged public sympathy had been shaped by continued media exposure to trolling victims:

- Trolling is not going away, but the public understands the dangers more because there has been much publicity about it. This is partly because there has been more research around trolling, and partly because there have been more stories about trolling.

Reflecting on the #StopTheTrolls campaign, Price linked external pressures from cumulative exposure to the combative action initiated by media organisations in Australia:

- I would have never imagined in a million years that a tabloid [The Daily Telegraph] would be conducting a campaign that said, let’s stop trolling, let’s stop online abuse of women. The public perception is very different than it was eight years ago.

In confirmation of trolling support strategies at Nine (formerly Fairfax Media) that owns the Canberra Times, that she writes a column for, Price stated:

- I was provided lots of information by my employer.

Describing the support provided by the newsroom, Price said colleagues were the most impactful resource:

- The most crucial support is none of the printed information, it is having a support person within the organisation who will help and support you if you are feeling down. I have found it extremely useful to have that person. The person who deals with my copy is also a fantastic support when the going gets tough. Every single person who writes about these things has a contact person, their supervisor.

When further questioned about the timeline of trolling support being introduced, Price explained that at Nine (formerly Fairfax Media), training strategies evolved
over time; however, work relationships that offered support were always present:

- I have always been able to speak to my editors about what is happening to me. Everyone has a direct chain where you can discuss these things with other people.

The importance of connections with family and friends was also raised. Price reiterated her colleagues outside of Nine (formerly Fairfax Media) were not only an alternative safety net, but a preferred one:

- Importantly, I also have a network of other women columnists where we talk about the things that are happening to us in a group chat.

When asked if newsrooms had edited the content of stories or removed her bylines to reduce personal trolling attacks on their employees, Price said:

- The answer is no. I am able to continue writing what I need to write and do not stop writing what I want to write, but I really take good care of my mental health.

When asked if personal resilience had grown over time, or was considered a fundamental and expected characteristic of journalism work, Price said:

- I have always been a pretty resilient person, which is why I have been able to continue as a journalist for a very long time.
- But I see young women who are less resilient and people who get off Twitter because they cannot stand the abuse.
- We need to all support each other.
- My own research discovered that the community of women who were the administrators and moderators really supported each other, and that made it possible for us to be resilient. It was fantastic.

Price raised the commonly advocated strategy of ‘don’t feed the trolls’ without criticism:

- In the early days, I’d be having discussions, trying to persuade them, sending them research but none of that works. But I don’t engage with those people anymore. I mute, I block or I delete.
• When trolling first started to happen to me I had no idea what was going on as the use of social media was new. Initially, you imagine you would be able to fix people, but it is absolutely not possible.

In summary and discussion, Price described how continuous media exposure engendered public awareness of the impacts of trolling, and subsequently framed public sympathy. An anti-trolling discourse put external pressure on newsrooms to respond in ways Price had never thought probable, including *The Daily Telegraph’s* #StopTheTrolls campaign. Continued pressure encouraged newsrooms to recognise trolling as a widespread workplace problem and acknowledge the potential for adverse psychological impacts on employees. This prompted the evolution of internal procedures for trolling-specific counselling and training resources, including the provision of support staff. Although *Fairfax Media* provided Price with ample information, colleagues were clearly identified as her preferred and most impactful resource, with the support of female colleagues both inside and outside the organisation being of particular help. This stands in contrast to comments by all 10 participants in chapter 4.2’s interviews who expressed they required more training and workplace support strategies than offered by their newsrooms.

Both Beckett and Price noted that the idea of reliance on resilience as a key coping strategy was outdated, and newsrooms needed to provide further support for employees in contemporary journalism where dealing with trolling has become a prominent part of the workplace. These views reflected the sentiments identified in the 2015 interviews with ten trolled journalists (in section 4.1 on p. 28). Participants highlighted that not only were there no universal workplace procedures in regard to trolling at the time of the interviews in 2015, but all felt this was inadequate. Their sense of a need for hyper-resilience may be an indication of a broken system, with the prolonged reliance on tolerance as coping strategy recognised as a band-aid solution to a widespread problem not comprehensively understood. Attitudes in both the 2015 victims interviews and 2020 academic interviews suggest further research to develop an understanding of best-practice is essential, to prepare journalists for a world in which vitriolic
discourse is increasingly prevalent.

Summary

The *ABC* was a global leader in the hard-line response to trolling. The removal of comment streams meant that no form of moderation was required. Other Australian newsrooms were much slower to respond, opting for user-generated content moderation methods for several years until the sporadic introduction of machine-learning software. Gradually, external pressure grew until consensus across all Australian newsrooms emerged in 2015, acknowledging trolling as a widespread problem in contemporary journalism. Two key events that prompted the most notable response from Australian newsrooms included:

- #StopTheTrolls campaign in 2012
- The Senate inquiry to access the *Adequacy of existing offences in the Commonwealth Criminal Code and of state and territory criminal laws to capture cyberbullying* in 2018.

Although multiple strategies have been adopted to varying degrees, the continued presence and virulence of trolling in the current media landscape suggests these approaches have not yet been successful in reducing abuse to a manageable level. The cost of cyberhate is not simply hurt feelings: it can result in job loss as well as poor mental health. In a world where journalists are required to have an online presence, trolling can be considered a form of economic vandalism. With no universal guidelines for journalists and newsrooms, further research is required to provide best-practices that also embrace freelance journalists. Gorman (2019) summed up current sentiments stating “while we have clear laws about the duty to maintain a safe workplace that are imposed on our employers, who is responsible for maintaining a safe social media space? That’s probably not just employers” (p. 71). As Ackoff’s (1974) System Theory suggests, an approach from multiple stakeholder groups is required if the social problem is to be adequately addressed. Although a fundamental portion, newsrooms are only a part of the complex and intricate.

This chapter has outlined ten responses to trolling by Australian newsrooms, providing clarity about an important stakeholder in the trolling paradigm. The
analysis of the relative success various newsroom initiatives highlights the importance of the support level of the systems model, a finding that resonates with comments by the ten trolled Australian journalists interviewed in chapter 4.2. However, without reinforcement from other levels, its impact is limited. The next chapter will investigate Australian legislative responses.
Chapter 4.7
Legal Responses

Results and Discussion

Criminal and civil law flowcharts for prosecuting a troll

*Question G – What Australian laws can be used by journalists to prosecute a troll?*

Current legislation is not effective for Australian journalists, who receive frequent abuse as a commonly accepted part of their job. As a legal grey area few trolls prosecuted and convicted in Australia. Trolling is a very broad church, ranging from amateur philosophers to bullies, from the mildly offensive to the illegal (Bartlett, 2013, p. 45), and the line between argument, criminality, threats, offensiveness, and satire is a very fine one (p 42). At present, there is no Australian law, statute or common, specific to trolling; however, criminal laws of sedition, libel and hate speech, and civil laws of harassment, stalking, defamation, discrimination, vilification, blasphemy, and obscenity can be drawn upon for prosecution (Pearson, 2019). However, legal addenda will take time, and victims during this time lag must deal with the life-changing consequences without compensation, and sometimes without adequate workplace support systems in place (Citron, 2014, p 226). Hateful and aggressive online communication is now commonly considered to be a social problem requiring legal interventions (Windisch, Wiedlitzka & Olaghere, 2021; Blaya, 2019). At the same time, online aggression remains a challenging phenomenon for authorities and public policy makers to manage (Kaakinen et al., 2018, p. 31). This chapter unpacks current laws and draws upon documented case studies in Australia to critique the legislative response, described by many victims as slow, unclear and ineffective.

The research design strategy that addresses Question G: *What Australian laws can be used by journalists to prosecute a troll?* applies mixed methods of archival evidence, thematic critical analysis and interviews. Although framing cyber-
bullying as a significant issue of the digital media space, the Australian government does not have a standalone trolling law. This chapter outlines although abuse can be covered by existing laws, the extent of harms and continued impacts on victims may indicate more direct laws are required to protect Australian journalists. In the media law textbook, The journalist’s guide to media law: A handbook for communicators in a digital world, Pearson and Polden (2019) described relevant Australian laws to prosecuting trolling. To provide primary data, the co-authors of this textbook Mark Pearson, an academic, media legal expert and former section editor of The Australian, and Mark Polden, an academic and media lawyer, were interviewed to extend on the content on their book. Both are Australian and were reached through academic and industry contacts. The semi-structured interview went for approximately one hour and provided both quantitative and qualitative data. Through predominantly open-ended questions, robust qualitative data was acquired. Developing patterns regarding the use and effectiveness of Australian laws in relation to trolling were drawn upon for comparison, discussion and critique. Through brainstorm discussions in two separate interviews with Pearson and Polden, a new list of general umbrella laws that may cover trolling was compiled. Each key area was then investigated, using the legal search engine www.austlii.edu.au. Thematic analysis was applied to categorise legal pathways as criminal or civil, and federal or state. The following websites were then searched to retrieve laws that could be used for prosecution, based on the victim’s jurisdiction:

- Northern Territory laws - www.legislation.nt.gov.au
- Tasmanian laws - www.legislation.tas.gov.au
- Western Australian laws - www.legislation.wa.gov.au

This complex process was difficult and time-consuming. For a trolled journalist
dealing with the emotional strain of the abuse, identifying the appropriate laws to apply would be a problematic and potentially unachievable task. In most cases this would lead to the trolled journalist having to engage a lawyer at their own expense (Whitbourn, 2022), further stressing their mental and financial well-being. Complicated legal pathways can also present issues for newsroom governance.

This study extends on Pearson and Polden’s published work, updating and presenting the specific Australian laws journalists are able to use against a troll in a useful format. This qualitative data is displayed in descriptive passages within themed tables to provide a comprehensive list for real-world implementation by trolled journalists. The descriptive passages of each law are notably long; however, each is directly quoted from their respective federal or state Criminal Codes and they are important for inclusion, to allow comparisons and critical analysis. Furthermore, this data can be used by future studies to assess the effectiveness of current trolling laws. From this data, two flowcharts, one criminal and one civil, were constructed to distil the complex dataset into a practical guide that trolled journalists can refer to when seeking to prosecute their abusers. The flowcharts are displayed first, to provide an overall context, followed by the data results for each theme and discussion. Criminal laws are described first, followed by civil laws.

**Brief background of legal jurisdictions in Australia**

Australia is a federation of states with a Westminster style legal and parliamentary system. Australia combines nine major jurisdictions including six states: New South Wales; Queensland; South Australia; Tasmania; Victoria; and Western Australia; and two territories: the Northern Territory and the Australian Capital Territory. The final jurisdiction is the Australian Commonwealth Government. As per the *Commonwealth of Australia Constitution 1900* (Cth), each state may compose laws on any subject matter, while the Commonwealth government can only make laws on areas that fall within the powers granted under the constitution. Where there is overlap, the states may still make laws, but those laws cannot contradict laws passed by the Commonwealth government. Criminal law is a breach of criminal codes prosecuted by the Commonwealth or State. Civil law
exists between private litigants where the issue is financial compensation.

**Brief background of trolling convictions in Australia**

On March 25, 2011, the first convicted Australian internet troll, Bradley Paul Hampson, was jailed for posting offensive pornographic material on Facebook tribute sites dedicated to slain Queensland children, 12-year-old Elliott Fletcher and 8-year-old Trinity Bates (Rawlins, 2011). Hampson was sentenced to three years jail, to serve 12 months of that sentence, after he pleaded guilty to one count each of distributing child exploitation material and possessing child exploitation material, and two counts of using a carriage service to menace, harass or cause offence (AAP, 2011). However, the Court of Appeal cut the actual time served from 12 months to six months after Hampson had served only 220 days, referencing Hampson’s limited criminal history and his autism, which may have hindered his insight into the impact of his offending (Rawlins, 2011). The case received widespread publicity in the Australian news media and set the precedent for future trolling prosecutions. Despite this case, the Australian government was criticised for minimal convictions resulting in jail time (Dwyer, 2016), including the widely publicised case of troll Zane Alchin who received a twelve-month good behaviour bond after posting sexually explicit and threatening comments on Facebook to Olivia Melville in 2015 about her Tinder profile.

**Scope of the chapter**

While there are no laws specifically about trolling journalists, certain Australian laws could be used to prosecute a troll, including intimate images and child pornography laws. This type of trolling was not raised in the 10 in-depth interviews with the trolled journalists in chapter 4.1 (on p. 30) and is considered to be of lower risk of occurring to practising journalists. These laws were therefore determined to fall beyond the scope of this study and were not included in the following descriptive analysis.

This study notes Australia administers a range of small non-mainland territories that include, but are not limited to, Ashmore and Cartier Islands, Australian Antarctic Territory, Christmas Island, Cocos (Keeling) Islands, Heard Island,
McDonald Islands and Norfolk Island. The study acknowledges these territories have their own respective laws, which were determined to exceed the scope of this study and were excluded from the data set.

### 4.7.1 Qualitative data results of interviews

Qualitative data was retrieved from in-depth interviews one hour in length via telephone, with academic and media law expert Professor Mark Pearson and academic and lawyer Mark Polden, who co-wrote *The journalist’s guide to media law* (2019), now in its 6th addition, and used as the textbook for journalism studies by several tertiary educators. Through thematic analysis of brainstormed discussions, a list of general umbrella laws that may cover trolling was compiled and presented in the table below.

#### Figure 32

*Australian umbrella laws for trolling.*

<table>
<thead>
<tr>
<th>Criminal Laws</th>
<th>Civil Laws</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Harassment</td>
<td>• Defamation</td>
</tr>
<tr>
<td>• Threats</td>
<td>• Personal injury</td>
</tr>
<tr>
<td>• Inciting suicide</td>
<td>• Worker compensation</td>
</tr>
<tr>
<td>• Vilification</td>
<td></td>
</tr>
<tr>
<td>• Stalking</td>
<td></td>
</tr>
<tr>
<td>• Criminal Defamation</td>
<td></td>
</tr>
<tr>
<td>• Obscenity</td>
<td></td>
</tr>
</tbody>
</table>

This list formed the starting point for the research from which the investigative process extended. The in-depth interview also provided qualitative data from discussion about the effectiveness of current Australian legislation in the contemporary journalism landscape, in which online interactivity with audiences plays a major role and has become a standard workplace practice.

#### Interview with Professor Mark Pearson

In November 2020, an interview was conducted with Professor Mark Pearson, which provided key insights on emerging problems with the practical application of Australian laws by journalists in response to frequent and malicious online
abuse. Responses were grouped into key themes that are described and summarised below.

Asked if it was possible to ascertain how many times the law “using a carriage service to menace, harass or offend” had been used successfully to prosecute a troll, Pearson said collecting and analysing such data would fall well beyond the scope of this thesis:

- This would be a huge task, a minefield.
- It would take a criminologist to be able to work out where records of particular prosecutions are kept.
- It would require access to police records in order to find the complaints, and then lower court records would need to be accessed to find out how many times people have been charged with these particular offenses.

The extensive criminology work required to retrieve relevant data with which to draw a quantitative conclusion was therefore not included as part of this study; however, specific laws were identified, and qualitative data was analysed, discussed and critiqued.

Reflecting on the use of criminal laws in practical application, Pearson highlighted that many laws are used in unison and, therefore, the true extent of trolling may not even be evident in a large-scale criminology study:

- Online abuse may often fall under a suite of charges.
- Many instances may be hidden under other laws that would include domestic violence orders.

Confirming the importance of this chapter that addresses the legal stakeholder group of the trolling paradigm, Pearson said:

- This is why it is so important to define prosecution routes.
- You will be exposing the fact that such information is so difficult to access, how it is not pulled together and how difficult it can be to do that without a legal background.

Spotlighting the need for enforcement officers to understand how to enforce
current legislation, Pearson said:

- Policing systems have a scale of importance for different crimes.
- Physical crimes are ones that historically the police and the prosecutors and the courts have been used to dealing with.
- Similarly, authorities need to find a way to collaborate and share information to prosecute criminals when they commit very serious crimes online.

Pearson described the difficulties that arise when prosecuting a troll across different jurisdictions, and identified the abuser’s location as a hindering factor that determined which legal avenues could be pursued:

- Online communications are ethereal and by definition cross-jurisdictional and are therefore more complicated than communication in the real world.
- An extradition order is required to bring the accused into the court of a neighbouring state.

Reflecting on his own experience of being trolled, Pearson criticised the practical application of current Australian laws in response to spiteful abuse online.

- I know from my own experience the process is difficult.
- I was called a rapist.
- I was going to make a complaint to the police, but they are very quick to refer it another state or territory.

Asked about the adequacy of social media companies removing reported trolling, Pearson described their response as underwhelming, but explained possible defences:

- Reasonableness is always an issue.
- Many platforms receive millions and millions of complaints.
- Removal of comments within two weeks may be an acceptable excuse for a smaller company; however, big platforms should be better staffed and ought to be able to remove offensive comments within 24 hours.

Asked if social media companies such as Facebook and Twitter should be
considered publishers and subsequently be held liable for the abusive commentary they host, Pearson said:

- Once they start to play a role in the publishing with the introduction of algorithms, they no longer remain secondary publisher.
- Innocent dissemination is the defence against defamation traditionally used by newsagents that sell newspapers. However, it is not applicable in this context.

In summary, Pearson described multiple problems with the application of current legislation and highlighted the need for further education and training about how police can implement these laws when digital abuse crims are drawn to their attention. Current legislation is not effective for Australian journalists, who receive frequent abuse as a commonly accepted part of their job. The legal stakeholder group that this chapter begins to address is a crucial part of a multi-layered approach needed to adequately respond to the social issue.

**Interview with barrister Mark Polden**

In October 2020, an interview was conducted with Mark Polden, which provided key insights into the inadequacy of current Australian laws that often fail to protect journalists from frequent and malicious online abuse and do not adequately reprimand trolls or the companies that host them. Responses were grouped into key themes that are described and summarised below.

Asked about how to approach social media platforms about removing abusive commentary, Polden said:

- For media purposes, in practical terms, what one would contact the platform and say this user is in breach of the terms of services. This user is breaching the contract with you, and it is affecting me. It ought to be taken out because it is breaching your terms of service.

When questioned about how to approach social media platforms to unmask a troll, Polden described the key steps but recognised further fundamental problems that can hinder the ability to reveal their true identity:
• First you would ask for the platform to identify the name of the user.
• If they refuse, you've taken reasonable steps to obtain that information.
• Then you would need to get an order for preliminary discovery to make them disclose the name or the URL or the identifying features of the person who posted.
• However, the person can hide their identity pretty well.
• If they are hiding behind some proxy server, their location can be bounced around the world.
• You can issue a subpoena to the American company under The Hague Convention. While it can be done, the federal court will often not issue the subpoena as a matter of discretion, because it is unenforceable. There is no power to enforce the laws in the US.
• With a lot of these entities, you run into the consistent problem that they stonewall courts in Australia because they are domiciled overseas.

Reflecting on whether social media platforms should be considered publishers after the Voller v Fairfax Media, Nationwide News and Sky News case in 2019, Polden said:

• Social media companies are responsible for the material; however, the Broadcasting Services Act provides immunity to Internet service providers in certain circumstances.
• However, if you are made aware of defamation or abuse on your site, then that immunity goes away.
• You would need to look at the terms of service of the platforms, which will be available, but I am reasonably confident that they would have provision there that would cover it.

On review of the terms and conditions of Facebook, the study identified this was addressed in section “3. Limits on liability” that stated:

We do not control or direct what people and others do or say, and we are not responsible for their actions or conduct (whether online or offline) or any content they share (including offensive, inappropriate, obscene, unlawful, and other objectionable content).
Twitter’s terms and conditions also addressed liability, in section “3. Content on the services,” which stated:

Any use or reliance on any content or materials posted via the services or obtained by you through the services is at your own risk. We do not endorse, support, represent or guarantee the completeness, truthfulness, accuracy, or reliability of any content or communications posted via the services or endorse any opinions expressed via the services. You understand that by using the services, you may be exposed to content that might be offensive, harmful, inaccurate or otherwise inappropriate, or in some cases, postings that have been mislabelled or are otherwise deceptive. All content is the sole responsibility of the person who originated such content. We may not monitor or control the content posted via the services and, we cannot take responsibility for such content.

(Twitter, 2021, para. 5)

Polden concluded:

- If the content is defamatory, then the social media platform is going to be liable for it.
- If social media companies dismiss liability, then pressure needs to be applied to parliamentarians to take them on.

Confirming the need for new laws against doxing, swatting, impersonation or cheap fakes that more adequately address online abuse in Australian newsrooms, Polden suggests continued political pressure from multiple stakeholder groups would be the most effective way to incite legislative change:

- [His] perception is that very little happens unless there is either a vested interest attached to it, or to seek political advantage.
- In order to get attention, it would need a concerted campaign.
- It would need schoolteachers worried about the effects on children
- It would need the clinical psychologists and Australian Psychological Association, the peak body for registered clinical psychologists.
- It would need long and concerted pressure, if things don't happen automatically.
- It puts it puts pressure on members of state, the Attorney-General, on the
Polden argued current Australian laws often fail to protect journalists from frequent and malicious online abuse that can impact the emotional and mental wellbeing of victims. Continued political pressure from multiple stakeholder groups is required to incite the government to consider specific law reform that holds trolls individually accountable, along with the platforms that fail to monitor and efficiently remove their abuse. This study acknowledges that this process is a lengthy one; however, the proposed new laws presented to parliament in 2021 (discussed later in this chapter), are acknowledged as a critical first step enforcing criminal penalties that reflect the severity of victim impacts.

### 4.7.2 Qualitative data results of flowcharts

**Flowchart of Australian criminal laws**

Through thematic analysis of qualitative data results, two critical requirements that must be met were determined. The first key consideration was giving victims the ability to discover the identity of the troll. Without knowing the true identity of the abuser, no legal pathway against the troll could proceed. Anonymity provided by the internet emerged as a significant problematic factor in prosecuting a troll. The study identified a potential civil case against the social media platform on which the trolling arose; however, existing cases were sporadic and had had limited success. Although platforms continue to deflect responsibility for hosting trolls in 2021, the subject is an area of widespread active debate and will be discussed in further detail later in this chapter.

The second key consideration was the location of the troll. If the troll is not located in Australia, limited avenues for prosecution exist. Difficulties also arise when the victim and abuser live in different states within Australia, as governing laws sometimes differ between jurisdictions, further compounding the complexity, and restricting the ability to provide universal legal pathways. Globalisation of the internet emerged as yet another problematic factor, due to the trans-national nature of social media companies that permit instantaneous potential abuse on a global scale.
These two key considerations were included in the original flowchart of Australian criminal laws displayed in Figure 33 over the three pages that follow.
Figure 33 Flowchart of criminal laws that trolled journalists could use derived from chapter 4.7 data.
<table>
<thead>
<tr>
<th>Vilification</th>
<th>Racial vilification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>STATE LAWS</td>
</tr>
<tr>
<td></td>
<td>ACT - Section 67A of the Criminal Code 2002 (ACT)</td>
</tr>
<tr>
<td></td>
<td>SA - Section 4 of the Racial Vilification Act 1996 (SA)</td>
</tr>
<tr>
<td></td>
<td>WA - Section 77 and Section 80A of the Criminal Code Act Compilation Act 1913 (WA)</td>
</tr>
<tr>
<td></td>
<td>There are no specific laws in NT</td>
</tr>
</tbody>
</table>

| Religious vilification | | |
|-----------------------|---------------------|
|                       | STATE LAWS          |
|                       | ACT - Section 67A of the Criminal Code 2002 (ACT) |
|                       | There are no specific laws in NSW, NT, SA and WA |

| Sexuality vilification | | |
|------------------------|---------------------|
|                        | STATE LAWS          |
|                        | ACT - Section 67A of the Criminal Code 2002 (ACT) |
|                        | There are no specific laws in NT, SA, VIC and WA |

| Gender identity vilification | | |
|-------------------------------|---------------------|
|                                | STATE LAWS          |
|                                | ACT - Section 67A of the Criminal Code 2002 (ACT) |
|                                | There are no specific laws in NT, SA, VIC and WA |

| HIV/AIDS vilification | | |
|----------------------|---------------------|
|                      | STATE LAWS          |
|                      | ACT - Section 67A of the Criminal Code 2002 (ACT) |
|                      | There are no specific laws in NT, Qld, SA, VIC and WA |

| Disability vilification | | |
|-------------------------|---------------------|
|                         | STATE LAWS          |
|                         | ACT - Section 67A of the Criminal Code 2002 (ACT) |
|                         | There are no specific laws in NSW, NT, SA, Qld, VIC and WA |

| Obscenity | Offensive language | |
|-----------|--------------------||
|           | STATE LAWS         |
|           | NSW - Section 4A of the Summary Offences Act 1988 (NSW) |
|           | NT - Section 53 of the Summary Offences Act 1923 (NT) |
|           | QLD - Section 6 of the Summary Offences Act 2005 (Qld) |
|           | SA - Section 32 of the Summary Offences Act 1913 (SA) |
|           | TAS - Section 12 of the Police Offences Act 1935 (Tas) |
|           | VIC - Section 17 of the Summary Offences Act 1966 (Vic) |
|           | No specific laws in ACT and WA |

| Indecent material | | |
|-------------------|---------------------|
|                    | STATE LAW           |
|                    | SA - Section 33 of the Summary Offences Act 1913 (SA) |
|                    | There are no specific laws in ACT, NSW, NT, Qld, TAS, VIC and WA |

| Offensive behaviour | | |
|---------------------|---------------------|
|                     | STATE LAWS          |
|                     | ACT - Section 392 of the Crimes Act 1900 (ACT) |
|                     | NSW - Section 4 of the Summary Offences Act 1988 (NSW) |
|                     | NT - Section 47 of the Summary Offences Act 1923 (NT) |
|                     | QLD - Section 217 of the Criminal Code Act 1889 (Qld) |
|                     | SA - Section 23 of the Summary Offences Act 1913 (SA) |
|                     | TAS - Section 21 of the Police Offences Act 1935 (Tas) |
|                     | VIC - Section 17 of the Summary Offences Act 1966 (Vic) |
|                     | WA - Section 202 of the Criminal Code Act Compilation Act 1913 (WA) |

| Disorderly conduct | | |
|--------------------|---------------------|
|                    | STATE LAWS          |
|                    | NT - Section 47 of the Summary Offences Act 1923 (NT) |
|                    | QLD - Section 6 of the Summary Offences Act 2005 (Qld) |
|                    | SA - Section 7 of the Summary Offences Act 1913 (SA) |
|                    | TAS - Section 33 of the Police Offences Act 1935 (Tas) |
|                    | VIC - Section 11A of the Summary Offences Act 1966 (Vic) |
|                    | WA - Section 74A of the Criminal Code Act Compilation Act 1913 (WA) |
|                    | There are no specific laws in ACT and NSW |

| Indecent exposure | | |
|-------------------|---------------------|
|                   | STATE LAWS          |
|                   | ACT - Section 392 of the Crimes Act 1900 (ACT) |
|                   | NSW - Section 5 of the Summary Offences Act 1888 (NSW) |
|                   | NT - Section 50 of the Summary Offences Act 1923 (NT) |
|                   | Qld - Section 6 in the Summary Offences Act 1894 (Qld) |
|                   | SA - Section 23 of the Summary Offences Act 1913 (SA) |
|                   | TAS - Section 21 of the Police Offences Act 1935 (Tas) |
|                   | VIC - Section 15 of the Summary Offences Act 1966 (Vic) |
|                   | WA - Section 204 of the Criminal Code Act Compilation Act 1913 (WA) |

---

Chapter 4.7 Legal responses results and discussion

278
Descriptive tables of Australian criminal laws

Through thematic analysis interviews with Professor Mark Pearson and Mark Polden, the study identified seven criminal law themes including harassment, threats, inciting suicide, vilification, stalking, criminal defamation and obscenity. In addition, three civil law themes were identified and will be discussed later in this chapter. Each criminal law theme was investigated on federal and state legislative websites, to determine which specific Australian laws could be used for prosecution under each individual category. The findings are presented in descriptive tables and are categorised by theme, then by jurisdiction. The study provides one trolling example for each criminal law theme. Each table is followed by a brief discussion to provide context for critical analysis.

Determining seriousness of crime

The average State maximum penalties, excluding Tasmania as an outlier, were calculated and coded to determine their graded level of seriousness. The study considered imprisonment to be far worse than a financial reprimand. With the abolition of capital punishment in Australia in 1985 (Australian Institute of Criminology, 1987) and the Crimes Legislation Amendment (Torture Prohibition and Death Penalty Abolition) Act 2010 blocking any state or territory from reintroducing it (National Library of Australia, 2017), life imprisonment has become the most severe sanction under the criminal law. To provide a solid base from which to access the ‘seriousness’ of an act, the study looked at the maximum penalties of other crimes under the Criminal Code Act Compilation Act 1913 (WA) that are displayed in the following table.
### Figure 34

**Graded level of seriousness for crimes in Western Australia derived from chapter 4.7 data**

<table>
<thead>
<tr>
<th>Crime</th>
<th>Criminal Code Act Compilation Act 1913 (WA)</th>
<th>Maximum penalty</th>
<th>Coded level of seriousness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Murder</td>
<td>Section 279</td>
<td>Life imprisonment</td>
<td>'Most serious' (Life imprisonment)</td>
</tr>
<tr>
<td>Attempted murder</td>
<td>Section 283</td>
<td>Life imprisonment</td>
<td></td>
</tr>
<tr>
<td>Manslaughter</td>
<td>Section 280</td>
<td>Life imprisonment</td>
<td></td>
</tr>
<tr>
<td>Aggravated burglary</td>
<td>Section 401</td>
<td>20 years</td>
<td>'Very serious' (15 + years imprisonment)</td>
</tr>
<tr>
<td>Aggravated sexual penetration without consent</td>
<td>Section 326</td>
<td>20 years</td>
<td></td>
</tr>
<tr>
<td>Sexual offences against a child under 13</td>
<td>Section 320</td>
<td>20 years</td>
<td>'Serious' (10-15 years imprisonment)</td>
</tr>
<tr>
<td>Sexual penetration without consent</td>
<td>Section 325</td>
<td>14 years</td>
<td></td>
</tr>
<tr>
<td>Grievous bodily harm</td>
<td>Section 297</td>
<td>10 years</td>
<td></td>
</tr>
<tr>
<td>Serious assault</td>
<td>Section 318</td>
<td>10 years</td>
<td></td>
</tr>
<tr>
<td>Assault causing bodily harm</td>
<td>Section 317</td>
<td>7 years</td>
<td>'Less serious' (5-10 years imprisonment)</td>
</tr>
<tr>
<td>Indecent assault</td>
<td>Section 323</td>
<td>5 years</td>
<td></td>
</tr>
<tr>
<td>Common assault</td>
<td>Section 313</td>
<td>3 years</td>
<td>'Moderate' (Less than 5 years)</td>
</tr>
<tr>
<td>Trespass</td>
<td>Section 70A</td>
<td>12 months and a fine of $12,000</td>
<td></td>
</tr>
</tbody>
</table>

The study used this method as a general guide for comparison and acknowledges it does not represent the exact level of seriousness attributed by Australian law.

**Standalone findings**

The study identified the relevant Australian Human Rights Commissioner for each theme that a trolled journalist could contact to lodge a complaint to examine a discrimination\(^5\) claim thorough an independent enquiry. The AHRC complaint

---

\(^5\) Discrimination happens when a person, or a group of people, is treated less favourably than another person or group because of their background or certain personal characteristics (Australian Human Rights Commission, 2021).
form can be downloaded on the https://humanrights.gov.au website and submitted by post to the Australian Human Rights Commission GPO Box 5218 Sydney 5218, by email to complaints@humanrights.gov.au or by fax to 02 9284 9611.

The study also found Tasmanian legislation, unlike the approach in other Australian jurisdictions, does not contain graduated penalties that would allow for a determination of the relative seriousness of each individual law. Obscenity legislation described in the Police Offences Act 1935 (Tas) did provide individual penalties which were recorded, but for all other criminal laws governed by the Criminal Code Act 1924 (Tas), section 389 provided a general clause stating, “the punishment for any crime shall be by imprisonment for 21 years, or by fine, or by both such punishments, and shall be such as the judge of the court of trial shall think fit in the circumstances of each particular case”. The generalised 21-year maximum penalty is therefore applied to all Tasmania laws in the Criminal Code Act 1924 (Tas); however, this does not mean that offenders who are sentenced for any offence under the Code will get the maximum penalty (Tasmania Sentencing Advisory Council, 2020). Courts, rather than the legislature, exercise discretion in prosecution, and sentences are rarely as long as 21 years for crimes other than murder and treason. The study notes the 21-year maximum penalty included in findings may not reflect the actual time period determined by the court when a troll is prosecuted and acknowledges its inclusion will skew the data set. Although presented in the descriptive tables, for quantitative comparison and discussion, the study selectively removed the Tasmanian 21-year maximum penalty as an outlier where appropriate.

The study found most jurisdictions have their own penalty unit value, which increased each year. This enables fines to be increased without the need for legislative amendments. The monetary value of each fine was calculated using the penalty units as of June 2021; these were as follows: Australian Capital Territory $160, New South Wales $110, Northern Territory $158, Queensland $133.45, Tasmania $172, Victoria $165.22, Western Australia $110. South Australia did not use a penalty points system and quoted the exact fine amount within the Criminal
Code. Determining the exact amount of these fines allowed for cross-referencing to access the severity of laws to between jurisdictions that did not include imprisonment time.

**Criminal laws for harassment**

Figure 35 describes the Australian federal criminal law “using a carrier service to menace, harass or cause offence” that governs the theme of harassment. This federal law can be used in all eight respective states or territories for prosecution and is the most referenced law in response to trolling. An example of trolling in breach of harassment, which included ongoing harassment as well as menacing and offensive online behaviour, under this law is also provided.

**Figure 35**

*Australian criminal laws for harassment that trolled journalists could use derived from chapter 4.7 data*

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
<th>Maximum Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federal</td>
<td>Using a carrier service to menace, harass or cause offence</td>
<td>Section 474.17 of the Criminal Code Act 1995 (Cth)</td>
<td>A person is guilty of an offence if: (a) the person uses a carriage service; and (b) the person does so in a way (whether by the method of use or the content of a communication, or both) that reasonable persons would regard as being, in all the circumstances, menacing, harassing or offensive.</td>
<td>Imprisonment for 3 years.</td>
</tr>
</tbody>
</table>

Section 46P of the *Human Rights and Equal Opportunity Commission Act 1986 (Cth)* allows journalists to make a complaint to the Human Rights Commissioner, a position held in 2021 by Edward Santow.

With a maximum penalty of 3 years imprisonment, the offense can attract a substantial punishment term, indicating that Australian law considers this type of harassment as a 'moderate' crime.
Discussion

Since the inclusion of the provision in 1995, academic discussions continue to debate the effectiveness of harassment prosecution in practical application. The law was designed to address frequent insulting phone calls rather than abuse delivered by forms of online communication when the law was first introduced. In more recent discussions about the adequacy of this particular law, in 2014, the Australian Law Reform Commission reported the Australian Federal Police saying the law was “more than adequate to facilitate prosecution” (Australian Law Reform Commission, 2014, para, 11). However, the onslaught of abusive content online in 2014, and its further escalated presence in 2021, may suggest otherwise. Although raising concerns about intensifying abuse, Google proposed the problem, “was not that laws do not exist, but rather that there was a general lack of awareness of the existing criminal and civil laws available” (para. 12). Anecdotal evidence, such as Australian journalist Osman Faruqi (2019), suggests knowledge about the applicability of trolling laws may be lacking, and legal pathways may need to be both more universally understood and actively pursued by journalists. This chapter begins to address the information gap; however, further education campaign extensions are necessary to reach the broader media community. While lengthy deliberations continue about the necessity of law reform, the flaws in the current legal system remain. Whether by means of new or amended legislation, the time lag until effective trolling laws are implemented continues to impact upon the mental wellbeing of victims.

Criminal laws for threats

The study identified four sub-categories encompassed by the criminal law theme of threats; these include threats to kill, threats to cause harm, hoax threat and threat to destroy property and are described in the four tables below. Quantitative analysis of the data is presented under each table, and a qualitative discussion follows. An example of trolling in violation of each respective law is also provided.
Figure 36

*Australian criminal laws for threats to kill that trolled journalists could use derived from chapter 4.7 data*

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
<th>Maximum Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federal</td>
<td>Using a carriage service to make a threat to kill</td>
<td>Section 474.17 of the Criminal Code Act 1995 (Cth)</td>
<td>A person (the first person) is guilty of an offence if: (a) the first person uses a carriage service to make to another person (the second person) a threat to kill the second person or a third person; and (b) the first person intends the second person to fear that the threat will be carried out</td>
<td>Imprisonment for 10 years.</td>
</tr>
<tr>
<td>ACT</td>
<td>Threats to kill</td>
<td>Section 30 of the Crimes Act 1900 (ACT)</td>
<td>A person makes a threat to another person to kill that other person or any third person (a) intending that other person to fear that the threat would be carried out; or (b) being reckless whether or not that other person would fear that the threat would be carried out; and The threat is made (c) without lawful excuse; and (d) in circumstances in which a reasonable person would fear that the threat would be carried out.</td>
<td>Imprisonment for 10 years.</td>
</tr>
<tr>
<td>NSW</td>
<td>Documents containing threats</td>
<td>Section 31 of the Crimes Act 1900 (NSW)</td>
<td>A person who intentionally or recklessly, and knowing its contents, sends or delivers, or directly or indirectly causes to be received, any document threatening to kill or inflict bodily harm on any person.</td>
<td>Imprisonment for 10 years.</td>
</tr>
<tr>
<td>NT</td>
<td>Threats to kill</td>
<td>Section 166 of the Criminal Code Act 1983 (NT)</td>
<td>Any person who, with intent to cause fear, makes, or causes any person to receive, a threat to kill any person which threat is of such a nature as to cause fear to any person of reasonable firmness and courage, is guilty of an offence.</td>
<td>Imprisonment for 7 years.</td>
</tr>
<tr>
<td>QLD</td>
<td>Threats to murder</td>
<td>Section 308 of the Criminal Code Act 1899 (Qld)</td>
<td>Any person who, knowing the contents thereof, directly, or indirectly causes any person to receive any document threatening to kill any person, is guilty of a crime.</td>
<td>Imprisonment for 7 years.</td>
</tr>
<tr>
<td>SA</td>
<td>Threats to kill</td>
<td>Section 19.1 of the Criminal Law Consolidation Act 1935 (SA)</td>
<td>A person is guilty of an offence who: (a) threatens, without lawful excuse, to kill or endanger the life of another; and (b) intends to arouse a fear that the threat will be, or is likely to be, carried out, or is recklessly indifferent as to whether such a fear is aroused.</td>
<td>Imprisonment for 10 years for a basic offence, or 12 years for an aggravated offence.</td>
</tr>
<tr>
<td>TAS</td>
<td>Threats to kill</td>
<td>Section 162 of the</td>
<td>Any person who, knowing the contents thereof, wilfully, and with intent thereby to</td>
<td>Imprisonment for 21 years.</td>
</tr>
</tbody>
</table>
Tasmania recorded the highest maximum penalty of 21 years imprisonment for all criminal laws. However, the study again recognises this generalised law is applied to all legislation and does not reflect a true representation of the actual penalty imposed by court when prosecuted for any particular law. The Tasmanian maximum penalty of 21 years is therefore excluded from the quantitative data analysis to ensure the findings are not inflated and positively skewed.

The highest maximum penalties of the other eight jurisdictions that specify graduated penalties for each law can be more fairly analysed as comparative data to determine the relative seriousness of each individual law. With Tasmania removed as an outlier, Western Australia had the highest maximum penalty of 14 years imprisonment for the offence if it was committed in circumstances of racial aggravation. For a standard offence, six jurisdictions had 10 years imprisonment, and two had 7 years imprisonment. With an average maximum penalty of 9.14 years imprisonment, the data suggests the offense can attract a significant punishment term and may indicate that Australian law considers threats to kill as a ‘less serious’ crime.
### Criminal Laws for a Threat to Cause Harm

**Trolling Example:** "May your bloodied head rest on the edge of an Isis blade" (West, 2015, para. 15)

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
<th>Maximum Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federal</td>
<td>Using a carriage service to make a threat to cause harm</td>
<td>Section 474.17 of the Criminal Code Act 1995 (Cth)</td>
<td>A person (the first person) is guilty of an offence if: (a) the first person uses a carriage service to make to another person (the second person) a threat to cause serious harm to the second person or a third person; and (b) the first person intends the second person to fear that the threat will be carried out.</td>
<td>Imprisonment for 7 years.</td>
</tr>
<tr>
<td>ACT</td>
<td>Threats to inflict grievous bodily harm</td>
<td>Section 31 of the Crimes Act 1900 (ACT)</td>
<td>A person makes a threat to another person to inflict grievous bodily harm on that other person or any third person (a) intending that other person to fear that the threat would be carried out; or (b) being reckless whether or not that other person would fear that the threat would be carried out; and The threat is made (a) without lawful excuse; and (b) in circumstances in which a reasonable person would fear that the threat would be carried out.</td>
<td>Imprisonment for 5 years.</td>
</tr>
<tr>
<td>NSW</td>
<td>Documents containing threats</td>
<td>Section 31 of the Crimes Act 1900 (NSW)</td>
<td>A person who intentionally or recklessly, and knowing its contents, sends or delivers, or directly or indirectly causes to be received, any document threatening to kill or inflict bodily harm on any person.</td>
<td>Imprisonment for 10 years.</td>
</tr>
<tr>
<td>NT</td>
<td>Threats to cause harm</td>
<td>Section 200 of the Criminal Code Act 1983 (NT)</td>
<td>Any person who threatens to do any injury, or cause any detriment, of any kind to another with intent to prevent or hinder that other person from doing any act that he is lawfully entitled to do, or with intent to compel him to do any act that he is lawfully entitled to abstain from doing, is guilty of an offence.</td>
<td>Imprisonment for 2 years.</td>
</tr>
<tr>
<td>QLD</td>
<td>Threatening violence</td>
<td>Section 75 of the Criminal Code Act 1899 (Qld)</td>
<td>Any person who: (a) with intent to intimidate or annoy any person, by words or conduct threatens to enter or damage a dwelling or other premise; or (b) with intent to alarm any person, discharges loaded firearms or does any other act that is likely to cause any person in the vicinity to fear bodily harm to any person or damage to property. commits a crime.</td>
<td>Imprisonment for 2 years.</td>
</tr>
</tbody>
</table>

---

*Figure 37*

**Australian criminal laws for threats to cause harm that trolled journalists could use**

*derived from chapter 4.7 data*
<table>
<thead>
<tr>
<th>State</th>
<th>Legal response</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td>Threats to cause harm</td>
</tr>
<tr>
<td>TAS</td>
<td>Causing a person to receive a letter threatening injury to property</td>
</tr>
<tr>
<td>VIC</td>
<td>Threats to inflict serious injury</td>
</tr>
<tr>
<td>WA</td>
<td>Threats to cause harm</td>
</tr>
</tbody>
</table>

With the Tasmanian 21-year maximum penalty removed as an outlier, New South Wales had the highest maximum penalty for a threat to cause harm of 10 years imprisonment, two jurisdictions had 7 years, three had 5 years, and a further two had 2 years imprisonment respectively. With an average maximum penalty of 5.38 years imprisonment, the data suggests the offense can attract a substantial punishment term and may indicate that Australian law considers threats to kill as a ‘less serious’ crime.
### Figure 38

Australian criminal laws for hoax threats that trolled journalists could use derived from chapter 4.7 data

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
<th>Maximum Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federal</td>
<td>Using a carriage service for a hoax threat</td>
<td>Section 474.16 of the Criminal Code Act 1995 (Cth)</td>
<td>A person is guilty of an offence if: (a) the person uses a carriage service to send a communication; and (b) the person does so with the intention of inducing a false belief that an explosive, or a dangerous or harmful substance or thing, has been or will be left in any place.</td>
<td>Imprisonment for 10 years.</td>
</tr>
<tr>
<td>ACT</td>
<td>Untrue representations</td>
<td>Section 122 of the Crimes Act 1900 (ACT)</td>
<td>A person who knowingly makes in any manner an untrue representation to any other person, being a representation that tends to give rise to apprehension for the safety of any person (including the person making the representation and the person to whom it is made) or property, or both, commits an offence.</td>
<td>Imprisonment for 5 years, 100 penalty units, or both.</td>
</tr>
<tr>
<td>ACT</td>
<td>Making false statements about contamination of goods with intent to cause public alarm or economic loss</td>
<td>Section 139 of the Crimes Act 1900 (ACT)</td>
<td>A person must not make a statement that the person believes to be false with the intention of: (a) inducing the person to whom the statement is made or others to believe that goods have been contaminated; and (b) in that way, either: (i) causing public alarm or anxiety; or (ii) causing economic loss through public awareness of the contamination, or the possibility of contamination.</td>
<td>Imprisonment for 10 years, 200 penalty units, or both.</td>
</tr>
<tr>
<td>NSW</td>
<td>Conveying false information that a person or property is in danger</td>
<td>Section 93Q of the Crimes Act 1900 (NSW)</td>
<td>A person who conveys information: (a) that the person knows to be false or misleading, and (b) that is likely to make the person to whom the information is conveyed fear for the safety of a person or of property, or both, is guilty of an offence.</td>
<td>Imprisonment for 5 years.</td>
</tr>
<tr>
<td>NSW</td>
<td>Making false statements concerning contamination of goods with intent to cause public alarm or economic loss</td>
<td>Section 93M of the Crimes Act 1900 (NSW)</td>
<td>A person who makes a statement that the person believes to be false: (a) with the intention of inducing the person to whom the statement is made or others to believe that goods have been contaminated, and (b) with the intention of thereby: (i) causing public alarm or anxiety, or (ii) causing economic loss through public awareness of the contamination, is guilty of a crime.</td>
<td>Imprisonment for 10 years.</td>
</tr>
<tr>
<td>Territory</td>
<td>Legal Code</td>
<td>Relevant Section</td>
<td>Offence Description</td>
<td>Maximum Imprisonment</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>------------------</td>
<td>---------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>NT</td>
<td>Criminal Code Act 1983 (NT)</td>
<td>Section 148D</td>
<td>A person who makes a statement that the person believes to be false: (a) with the intention of inducing the person to whom the statement is made or others to believe that goods have been contaminated; and (b) with the intention of thereby: (i) causing public alarm or anxiety; or (ii) causing economic loss through public awareness of the contamination, is guilty of an offence.</td>
<td>Imprisonment for 10 years.</td>
</tr>
<tr>
<td>QLD</td>
<td>Criminal Code Act 1899 (Qld)</td>
<td>Section 239</td>
<td>A person who makes a statement or conveys information to another person that he or she knows or believes to be false with the intention of inducing in that person or another person a belief that goods have been contaminated or interfered with and causes: (a) public alarm or anxiety; or (b) that person or that other person to refrain from purchasing those goods or goods of that or any similar class; or (c) any person to suffer economic loss through taking steps to avoid public alarm or anxiety; or (d) members of the public to refrain from purchasing those goods or goods of that or any similar class; commits a crime.</td>
<td>Imprisonment for 7 years.</td>
</tr>
<tr>
<td>SA</td>
<td>Criminal Law Consolidation Act 1935 (SA)</td>
<td>Section 261</td>
<td>A person is guilty of an offence if the person: (a) contaminates goods; or (b) makes it appear that goods have been, or are about to be contaminated; or (c) threatens to contaminate goods; or (d) falsely claims that goods have been or are about to be contaminated, intending (e) to influence the public against purchasing the goods or goods of the relevant class or to create an apprehension that the public will be so influenced; and (f) by doing so: (i) to gain a benefit for himself, herself, or another; or (ii) to cause loss or harm to another.</td>
<td>Imprisonment for 5 years.</td>
</tr>
<tr>
<td>TAS</td>
<td>Criminal Code Act 1924 (Tas)</td>
<td>Section 276AA</td>
<td>A person who makes a statement or conveys information, being a statement or information that he knows to be false, to the effect that, or from which it could be reasonably inferred that, some act has been, will be, or is likely to be, done at any place that is of such a nature as to give rise, or be likely to give rise, to serious risk of danger to persons or property at or near that place, is guilty of a crime.</td>
<td>Imprisonment for 21 years.</td>
</tr>
<tr>
<td>TAS</td>
<td>Criminal Code Act 1924 (Tas)</td>
<td>Section 287E</td>
<td>A person who makes a threat that goods will be contaminated with the intention of: (a) causing public alarm or anxiety; or (b) causing economic loss through public awareness of the contamination, is guilty of an offence.</td>
<td>Imprisonment for 21 years.</td>
</tr>
<tr>
<td>Jurisdiction</td>
<td>Description</td>
<td>Act/Section</td>
<td>Punishment</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
<td>-------------</td>
<td>------------</td>
<td></td>
</tr>
<tr>
<td>VIC</td>
<td>Making false statements concerning contamination of goods with intent to cause, or being reckless as to whether it would cause, public alarm or economic loss</td>
<td>Code Act 1924 (Tas) Section 251 of the Crimes Act 1958 (Vic)</td>
<td>A person who, without lawful excuse, makes to another person a threat to inflict serious injury on that other person or any other person: (a) intending that that other person would fear the threat would be carried out; or (b) being reckless as to whether or not that other person would fear the threat would be carried out is guilty of a crime. Imprisonment for 10 years, 1200 penalty units, or both.</td>
<td></td>
</tr>
<tr>
<td>WA</td>
<td>Statement or act creating false apprehension as to existence of threat or danger</td>
<td>Section 388C of the Criminal Code Act Compilation Act 1913 (WA)</td>
<td>Any person who makes a statement or conveys information which that person knows to be false, and which expressly indicates, or may reasonably be construed as indicating (a) that an unlawful threat has been made; or (b) that there has been, is, or is to be an intention, proposal, plan, or conspiracy to unlawfully (i) kill, injure, endanger, or harm any person, whether a particular person or not. (ii) destroy, damage, endanger or harm any property, whether particular property or not; or (iii) take or exercise control of a building, structure or conveyance by force or violence; or is guilty of a crime. Imprisonment for 10 years.</td>
<td></td>
</tr>
</tbody>
</table>

The Australian Capital Territory and New South Wales have two criminal laws in relation to a hoax threat. For this analysis, the study assumes the troll published the abusive comments with the intent to cause alarm, and therefore the longer maximum penalties are included for comparison. With the Tasmanian 21-year maximum penalty removed as an outlier, six jurisdictions had the highest maximum penalty for a hoax threat of 10 years imprisonment, one had 7 years and one had 5 years. With an average maximum penalty of 9 years imprisonment, the data suggests the offense can attract a substantial punishment term and may indicate that Australian law considers hoax threats a ‘less serious’ crime.
### Criminal Laws for a Threat to Destroy Property

**Trolling Example:** "I'll burn your fucking house down to the ground" (West, 2015a, p. 77)

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
<th>Maximum Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>Threat to cause property damage</td>
<td>Section 407</td>
<td>A person commits an offence if the person: (a) intentionally makes to someone else a threat to damage property belonging to that person or another person; and (b) intends that person to fear that the threat will be carried out.</td>
<td>Imprisonment for 2 years, 200 penalty units (equivalent to $32,000 as of June 2021) or both.</td>
</tr>
<tr>
<td>NSW</td>
<td>Threatening to destroy or damage property</td>
<td>Section 199</td>
<td>A person who, without lawful excuse, makes a threat to another, with the intention of causing that other to fear that the threat would be carried out—(a) to destroy or damage property belonging to that other or to a third person, or (b) to destroy or damage the first-mentioned person’s own property in a way which that person knows will or is likely to endanger the life of, or to cause bodily injury to, that other or a third person.</td>
<td>Imprisonment for 5 years.</td>
</tr>
<tr>
<td>NT</td>
<td>Threat to damage property</td>
<td>Section 241</td>
<td>A person is guilty of an offence if the person makes a threat to another person to cause damage to property belonging to that other person or someone else.</td>
<td>Imprisonment for 2 years, or 7 years imprisonment if carrying out the threat will cause death or serious harm to someone.</td>
</tr>
<tr>
<td>QLD</td>
<td>Sending letters threatening to burn or destroy</td>
<td>Section 478</td>
<td>Any person who, knowing the contents of the document, causes any person to receive any document threatening that any building or vessel, whether complete or not, or any stack of cultivated vegetable produce, or any such produce that is in or under a building, shall be burnt or destroyed, is guilty of a crime.</td>
<td>Imprisonment for 7 years.</td>
</tr>
<tr>
<td>SA</td>
<td>Threats to damage property</td>
<td>Section 85</td>
<td>A person who, without lawful excuse, threatens to damage another’s property (a) intending to arouse a fear that the threat will be, or is likely to be, carried out; or (b) being recklessly indifferent as to whether such a fear is aroused, is guilty of an offence.</td>
<td>Imprisonment for 5 years for a basic offence, 7 years for an aggravated offence, or 15 years imprisonment for an aggravated offence by</td>
</tr>
<tr>
<td>Jurisdiction</td>
<td>Description</td>
<td>Relevant Section</td>
<td>Summary</td>
<td>Penalty</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
<td>-----------------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>TAS</td>
<td>Sending letters threatening to burn or destroy</td>
<td>Section 276 of the Criminal Code Act 1924 (Tas)</td>
<td>Any person who, knowing the contents thereof, directly, or indirectly causes any person to receive any writing threatening that any property shall be unlawfully burnt, destroyed, or injured, is guilty of a crime.</td>
<td>Imprisonment for 21 years.</td>
</tr>
<tr>
<td>TAS</td>
<td>False threats of danger</td>
<td>Section 276AA of the Criminal Code Act 1924 (Tas)</td>
<td>A person who makes a statement or conveys information, being a statement or information that he knows to be false, to the effect that, or from which it could be reasonably inferred that, some act has been, will be, or is likely to be, done at any place that is of such a nature as to give rise, or be likely to give rise, to serious risk of danger to persons or property at or near that place, is guilty of a crime.</td>
<td>Imprisonment for 21 years.</td>
</tr>
<tr>
<td>VIC</td>
<td>Threats to destroy or damage property</td>
<td>Section 198 of the Crimes Act 1958 (Vic)</td>
<td>A person who without lawful excuse makes to another a threat: (a) to destroy or damage any property belonging to that other or a third person or to himself and that other or a third person; or (b) to destroy or damage his own property in a way which he knows or believes is more likely than not to endanger the life of that other or a third person. If he made the threat with the purpose of causing the other to fear that it would be carried out, shall be guilty of an indictable offence.</td>
<td>Imprisonment for 5 years.</td>
</tr>
<tr>
<td>WA</td>
<td>Threat toward dwelling</td>
<td>Section 74 of the Criminal Code Act Compilation Act 1913 (WA)</td>
<td>Any person: (a) with intent to intimidate or annoy any person, threatens to enter, or damage a dwelling; or (b) with intent to alarm any person in a dwelling, discharges loaded firearms or commits any other breach of the peace; is guilty of a crime.</td>
<td>Imprisonment for 3 years. For a summary conviction imprisonment for 1 year and a $12,000 fine.</td>
</tr>
</tbody>
</table>

There are no specific federal laws for a threat to destroy property.

With the Tasmanian 21-year maximum penalty removed as an outlier, Queensland had the highest maximum penalty for a basic offence for a threat destroy property of 7 years imprisonment, three jurisdictions had 5 years, one had 3 years, and a further two had 2 years imprisonment respectively. With an average penalty of 4.14 years imprisonment, the data suggests the offense can attract a substantial punishment term and may indicate that Australian law considers threats to kill a ‘moderate’ crime.

Fuelled by anonymity, online trolls have emerged as one of the unfortunate by-products of the liberating impact of social media. An online culture has emerged
where vicious and frequent trolling has become commonplace and “I’m going to rape you” has become a predictable default response to journalism practice (Bartlett, 2013; Greig, 2016; Phillips, 2015; Citron, 2014; West, 2015; Williams, 2016; Tuohy, 2019; Hyland, 2021). Trolling terminology consists of offensive threats of rape and death, with explicit comments suggesting victims deserved to be urinated on and sexually abused (O’Connor, 2015). Although initial perceptions about the impacts of trolling were downplayed suggesting that, “unlike real rape, words and images on a screen cannot really hurt anyone” (Citron, 2014, p. 73). Numerous documented case studies are now published in academic data and provide countering evidence of tangible emotional and mental distress (Binns, 2017; Bossio & Holton, 2021; Gardiner, 2018; Lewis, Zamith & Coddington, 2020; Martin & Murrell, 2020; Martin & Murrell, 2021). These verified impacts are amplified by difficulty of navigating legislation.

One example that highlighted the additional trauma inflicted by the challenging and prolonged legislative process involved Australian journalist, Van Badham, who was trolled by far-right extremist, and former journalist Nathan Sykes. Although the magistrate court ruled in favour, reflecting on the targeted harassment campaign, Badham (as cited in Gorman, 2019) expressed serious concerns about the process saying:

> It was a year of my life. There were six or seven court appearances, and I was strung out, exhausted, stressed out of my mind. I’m not a suspicious person by nature, but the ongoing cyberhate has changed my personality. I mourn the persona I used to be.

(p. 72)

The grave impact on her mental wellbeing was evident. Similar concerns voiced in the seminal case study Gamergate suggest a lack of legislative support can compromise the emotional and psychological welfare of victims and should be held partially responsible. In this prominent example, video game developer Brianna Wu has spoken out multiple times about how traumatic and exhausting the harassment of her family has been (Herzog, 2015). Yet despite widespread publicly, US-based Wu was unable to prosecute her abusers as the Federal Bureau of Investigation in the United States failed to identify the online trolls despite Google and Microsoft tracking and providing their IP addresses to assist the investigation (Dring, 2017). Providing insights about further problems with the
policing of harassment laws, former FBI special agent for cybercrimes Tim Ryan explained that cyber-abuse cases are a low priority for authorities as other crimes they are tasked to enforce attract higher penalties (Ryan, as cited in Hess, 2014). Similarly in Australia, the lack of continuity between applicable legislation, and the law enforcement agencies that are required to apply that legislation, has emerged as a significant obstruction to justice for trolling victims. While some journalists continue to receive daily harassment, continued problems with the judicial cycle permit trolls to sustain these activities with minimal risk of punishment.

**Criminal laws for inciting suicide**

Figure 40 below displays the Australian criminal laws for inciting suicide that could potentially be applied to Australian journalists such as Van Badham (Larsen, 2014), Ford (2016) and Dawson (Connelly & Keene, 2012) who received “go kill yourself” comments. There are no specific federal laws. If the victim attempted suicide as a result of trolling, prosecution is viable in all seven states. Tasmania is the only state in which the act of attempting suicide is not a requirement for prosecution. An example of comments that encourage suicide is also provided.

**Figure 40**

*Australian criminal laws for inciting suicide that trolled journalists could use derived from chapter 4.7 data*

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
<th>Maximum Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>Suicide aiding</td>
<td>Section 17 of the Crimes Act 1900 (ACT)</td>
<td>A person is guilty of an offence if the person: (a) incites or counsels another person to commit suicide; and (b) the other person commits, or attempts to commit, suicide as a consequence of that incitement or counselling.</td>
<td>Imprisonment for 10 years.</td>
</tr>
<tr>
<td>NSW</td>
<td>Suicide aiding</td>
<td>Section 31C of the Crimes Act 1900 (NSW)</td>
<td>A person is guilty of an offence if the person: (a) a person incites or counsels another person to commit suicide, and (b) that other person commits, or attempts to commit, suicide as a consequence of that incitement or counsel.</td>
<td>Imprisonment for 5 years.</td>
</tr>
</tbody>
</table>
Legal responses results and discussion

| NT      | Encouraging suicide | Section 162 of the Criminal Code Act 1983 (NT) | A person is guilty of an offence if the person:
|         |                    |                                           | (a) assists another person to kill or attempt to kill himself or herself
|         |                    |                                           | (b) encourages another person to kill or attempt to kill himself or herself.
|         |                    |                                           | For a person to be guilty of an offence:
|         |                    |                                           | (a) the person must have intended his or her conduct would encourage the other person to commit suicide; and
|         |                    |                                           | (b) the other person commits or attempts to commit suicide and was encouraged to do so by that conduct.
|         |                    |                                           | Imprisonment for life.
| QLD     | Aiding suicide     | Section 311 of the Criminal Code Act 1899 (Qld) | Any person who:
|         |                    |                                           | (a) procures another to kill himself or herself; or
|         |                    |                                           | (b) counsels another to kill himself or herself and thereby induces the other person to do so is guilty of a crime.
|         |                    |                                           | Imprisonment for life.
| SA      | Inciting suicide   | Section 13A of the Criminal Law Consolidation Act 1935 (SA) | A person who aids, abets or counsels the suicide of another, or an attempt by another to commit suicide, shall be guilty of an indictable offence.
|         |                    |                                           | Imprisonment for 8 years for attempted suicide. Imprisonment for 14 years for suicide.
| TAS     | Incitement suicide | Section 163 of the Criminal Code Act 1924 (Tas) | Any person who instigates or aids another to kill himself is guilty of a crime.
|         |                    |                                           | Imprisonment for 21 years.
| VIC     | Inciting suicide   | Section 6B of the Crimes Act 1924 (Vic) | Any person who incites any other person to commit suicide and that other person commits or attempts to commit suicide in consequence thereof.
|         |                    |                                           | Imprisonment for 5 years.
| WA      | Procuring suicide  | Section 288 of the Criminal Code Act Compilation Act 1913 (WA) | Any person who:
|         |                    |                                           | (a) procures another to kill himself; or
|         |                    |                                           | (b) counsels another to kill himself and thereby induces him to do so is guilty of a crime.
|         |                    |                                           | Imprisonment for life.

There are no specific federal laws for inciting suicide.

With the longest maximum penalty of all criminal laws, incitement of suicide carries a life imprisonment sentence in the three jurisdictions of the Northern Territory, Queensland, and Western Australia. With the possibility of life imprisonment, the data may suggest incitement of suicide is considered the ‘most serious’ form of trolling by Australian law; however, the broad range of maximum penalties between five years and life may imply a national indecisiveness about the seriousness of the behaviour. In most jurisdictions, only living victims can sue; however, the death of a victim does not impact upon the ability for the deceased person’s estate to sue. Tasmania is the only state in which trolls can be held liable
for inciting suicide for behaviours that do not lead to the self-harm or fatality of the victim.

Unlike discrepancies about the severity of inciting suicide laws, there appears to be a more general consensus about the seriousness of manslaughter. Deakin criminal law researchers advocate offenders should be charged with manslaughter if they exploit another person’s vulnerable mental state and cause them to commit suicide (McMahon & McGorrery, 2018). The suggested change would mean offenders could face longer jail time without parole and prosecutors may more easily be able to prove guilt. However, the complexity of this law must consider issues of free will, causation and shared responsibility. Dr McMahon (2017) explained:

a fundamental assumption of common law-based legal systems is that individuals have free will and are responsible for their own actions, which is why they are held liable for their behaviour when it constitutes a criminal offence. However, there are certain circumstances where the legal system will accept on person’s free will has been so overborne that responsibility is transposed onto another person.

(para. 10)

The broad range of maximum penalties highlight a national indecisiveness about the seriousness of this particular type of trolling. Academic debate is exploring the contributing role that online abuse plays in the complex interconnected web of suicide causation. Gorman (2019a) suggested cyber hate alone is unlikely to cause suicide; however, predator trolling can be a significant stressor in a person’s life (para. 2), particularly for journalists who are often required to frequently engage online as part of their work. The true danger lies in the impacts of trolling combined with other difficulties an individual may be experiencing at any given time. Badham (as cited in Gorman, 2019) further highlights the compounding struggles of those who are already grappling with depression, stating, “trolling fuels the worst instincts of your illness” (p. 68). The impact of underlying mental health issues may leave journalists vulnerable to trolling potentially emerging as a catalyst for suicide. While the sole blame cannot be attributed to the troll, this is not considered an acceptable defence for a reduced sentence and the act receives the highest maximum penalty of life imprisonment in three jurisdictions. The
study notes that the intense impact of suicide is far reaching, and it affects multiple stakeholders beyond the deceased person including partners, family, close friends, colleagues, and extended acquaintances. This widespread personal impact is a key determining factor in the study's suggestion to assign more serious penalties to this law. There is therefore a ground for the remaining six jurisdictions (the Australian Capital Territory, Northern Territory, South Australia, Tasmania, Victoria, and the Commonwealth) to consider adopting a life imprisonment sentence for inciting suicide through online abuse. A stronger and united stance may act as a more robust deterrence strategy.

Discussions about legal reform to adequately address the escalation of cyber abuse have circulated since news began the transition online (Windisch, Wiedlitzka & Olaghere, 2021; Blaya, 2019). However, these concerns did not attract state or Federal parliamentary debate until the catalyst event on January 3, 2018, in which 14-year-old Amy “Dolly” Everett took her own life after sustained cyberbullying. As the Australian ambassador for the outback hat company Akruna Everett suffered a targeted hate campaign, and reports of her suicide made international headlines (Arena, 2018). Dolly became the recognised face of trolling’s greatest potential impact, the death of a child, and forced the government to prioritise a response. Funded research led to the implementation of “Dolly’s Law” by the NSW Parliament, increasing the maximum penalty from 3 years to 5 years imprisonment. Continued public outrage about the exploitation of children online has resulted in repeated Federal legislation revision; however, the study questions the adequacy of such a slow legislative response, and highlights that minimal action was undertaken to proactively address trolling until Dolly’s death.

Criminal laws for criminal defamation

Defamation is part of tort law and allows a journalist to file a lawsuit over a civil wrong that has been done to them, such as trolling. Internationally, this area of law is often referred to as libel when defamation material is published or slander when words are spoken; however, uniform national defamation laws in Australian introduced in 2005 encompass both terms under the one law of defamation.
Although criminal defamation still exists in many liberal democracies, including Australia, prosecutions are rare and often unsuccessful (Pearson & Polden, 2019). These rare cases involve severe incidents by a perpetrator who lacks the money to pay for the associated damages that maliciously attack a journalist’s character. Despite its lower prosecution rate, the law can still be considered as a potential prosecution strategy for trolled journalists and was therefore determined as relevant for inclusion. Successful prosecutions award the victim with a sum of damages and may also impose a permanent injunction that prevents any further publication, or a take-down order that requires offensive comments to be removed. Although none are specified at federal level, criminal defamation laws are defined in the following table and an example of trolling is also provided.

**Figure 41**

*Australian criminal laws for criminal defamation that trolled journalists could use*

*derived from chapter 4.7 data*

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
<th>Maximum Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ACT</strong></td>
<td>Malicious publication of defamatory libel</td>
<td>Section 29 of the <em>Defamation Act 2001 (ACT)</em></td>
<td>A person must not maliciously publish a defamatory libel (whether or not the person knows it is false).</td>
<td>Imprisonment for 10 years.</td>
</tr>
<tr>
<td><strong>NSW</strong></td>
<td>Criminal defamation</td>
<td>Section 529(3) of the <em>Crimes Act 1900 (NSW)</em></td>
<td>A person who, without lawful excuse, publishes matter defamatory of another living person (the victim): (a) knowing the matter to be false, and (b) with intent to cause serious harm to the victim or any other person or being reckless as to whether such harm is caused, is guilty of an offence.</td>
<td>Imprisonment for 5 years.</td>
</tr>
<tr>
<td><strong>NT</strong></td>
<td>Criminal defamation</td>
<td>Section 204 of the <em>Criminal Code Act 1983 (NT)</em></td>
<td>Any person who unlawfully publishes any defamatory matter: (a) with intent to cause or that causes or is likely to cause a breach of the peace (b) with intent to cause loss. (c) with intent to interfere with the free and informed exercise of a political right. (d) with intent to prevent or deter a person from performing any duty imposed on him by law. (e) with intent to prevent or deter any person from doing any act that he is lawfully entitled to do or to compel him to do any act that he is lawfully entitled to abstain from doing.</td>
<td>Imprisonment for 3 years.</td>
</tr>
</tbody>
</table>
(f) with intent to prevent any lawful investigation or inquiry or
(g) with intent to interfere with or to influence any judicial proceedings.

| QLD | Criminal defamation | Section 365 of the Defamation Act 2005 (Qld) | Any person who, without lawful excuse, publishes matter defamatory of another living person (the "relevant person")
(a) knowing the matter to be false or without having regard to whether the matter is true or false; and
(b) intending to cause serious harm to the relevant person or any other person or without having regard to whether serious harm to the relevant person or any other person is caused commits a misdemeanour. | Imprisonment for 3 years. |
| SA | Criminal defamation | Section 257 of the Criminal Law Consolidation Act 1935 (SA) | A person who, without lawful excuse, publishes defamatory matter concerning another living person:
(a) knowing the matter to be false or being recklessly indifferent as to whether the matter is true or false; and
(b) intending to cause serious harm or being recklessly indifferent as to whether the publication of the defamatory matter will cause serious harm, to a person (whether the person defamed or not), is guilty of an offence. | Imprisonment for 3 years. |
| TAS | Criminal defamation | Section 196 of the Criminal Code Act 1924 (Tas) | A person who, without lawful excuse, publishes matter defamatory of another living person (the victim):
(a) knowing the matter to be false or without having regard to whether the matter is true or false; and
(b) intending to cause serious harm to the victim or any other person or without having regard to whether such harm is caused is guilty of a crime. | Imprisonment for 21 years. |
| VIC | Criminal defamation | Section 10 of the Wrongs Act 1958 (Vic) | Every person who maliciously publishes any defamatory libel knowing the same to be false shall be liable to imprisonment for a term of not more than two years and to pay such fine as the court awards. | Imprisonment for 2 years and a fine determined by the court. |
| WA | Criminal defamation | Section 365 of the Defamation Act 2005 (WA) | A person who, without lawful excuse, publishes matter defamatory of another living person (the victim):
(a) knowing the matter to be false or without having regard to whether the matter is true or false; and
(b) intending to cause serious harm to the victim or any other person or without having regard to whether such harm is caused is guilty of a crime. | Imprisonment for 3 years. |

There are no specific federal laws for criminal defamation.

With the Tasmanian 21-year maximum penalty removed as an outlier, the Australian Capital Territory had the highest maximum penalty of 7 years imprisonment for criminal defamation. One jurisdiction had 5 years, four had 3 years and one had 2 years imprisonment respectively. With an average maximum penalty of 4.14 years imprisonment, the data suggests the offense can attract a substantial punishment term and may indicate that Australian law considers
criminal defamation a ‘moderate’ crime. Successful prosecutions award the victim with a sum of damages and judges may also impose a permanent injunction that prevents any further publication or a take-down order that requires offensive comments to be removed.

Criminal defamation is based on the notion that for serious incidents there should be some means under the criminal law by which punishment of the offender can be secured. For these severe offenses, a troll can be prosecuted by the state itself and sentenced to jail in addition to financial reprimands (Pearson & Polden, 2019). While prosecutions are rare, it is important for journalists to know of its existence; hence its inclusion in the Pearson and Polden (2019) textbook. However, criminal defamation, along with seditious libel, the term used for defamation of the state, can be leveraged as censorship tools against the media by corporations, corrupt politicians, and government officials to silence dissidents.

Journalists have been generally opposed to defamation laws as they are often used against news organisations; but faced with spiteful threats of violence and repeated attempts to undermine professional credibility, some have begun to use this legislation to protect themselves against online harassment. South African journalist Anton Harber (as cited in Schiggrin, 2020) eloquently expressed these sentiments in the following statement:

> As a journalist I have deep aversion to defamation litigation. We want the freedom to say strong things, even push the bounds of acceptable speech, and not have defamation law used to silence us (however) we only embarked on this civil action because we could see no other way to stop things being said that were hurtful and harmful, even dangerous. (para. 8)

The study acknowledges that free expression is a critical requirement of robust journalism practise; however, without criminal code provisions many countries lack adequate protections for journalists facing online harassment and require journalists to use defamation legislation as a defensive measure. An advocate for this approach, Peruvian investigative journalist Gustavo Gorriti explained that commencing legal action was often enough for defamers to issue retractions. He encouraged journalists to take a more determined approach that utilised current
legislation as a protective strategy (Gorriti, as cited in Schiffrin, 2020). The deterrent effect of the law may teach or remind trolls about penalties. Although remedies include compensation for victims, compensation is often linked to lost income and suffering can be hard to quantify. The presence of hate on social media and the seeming inability to regulate the trend suggests criminal defamation may still have a valuable role to play.

Criminal laws for stalking

Stalking may be applied to online behaviour and is an illegal offence in seven states. Queensland is excluded from this group labelling the act unlawful, meaning the conduct is prohibited by the law but holds no criminal penalty. In an important note on the fundamental difference, Professor Simon Rice, former director of law reform and social justice at the Australian National University, stated:

Unlawful acts are pursued by the person or entity who is aggrieved, and illegal acts are pursued by the police in order to punish the perpetrator. For unlawful conduct the harmed person seeks a personal remedy such as compensation; for illegal conduct the perpetrator is punished (with imprisonment)

(Rice, as cited in ABC, 2014, para. 18).

Western Australia has two separate offences for stalking; however, there are none specified at federal level. An example of trolling in breach of the stalking law is also provided.

Figure 42

Australian criminal laws for stalking that trolled journalists could use derived from chapter 4.7 data

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
<th>Maximum Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>Stalking</td>
<td>Section 35 of the Crimes Act 1900 (ACT)</td>
<td>A person must not stalk someone with intent: (a) to cause apprehension, or fear of harm, in the person stalked or someone else; or (b) to cause harm to the person stalked or someone else; or</td>
<td>Imprisonment for 2 years.</td>
</tr>
</tbody>
</table>
(c) to harass the person stalked. For this section, a person stalks someone else (the stalked person) if, on at least 2 occasions, the person does 1 or more of the following:
(e) gives or sends offensive material to the stalked person or leaves offensive material where it is likely to be found by, given to, or brought to the attention of, the stalked person
(f) telephones, sends electronic messages to, or otherwise contacts the stalked person
(g) sends electronic messages about the stalked person to anybody else
(h) makes electronic messages about the stalked person available to anybody else
(i) engages in conduct amounting to intimidation, harassment, or molestation of the stalked person.

NSW Stalking Section 13 of the Crimes (Domestic and Personal Violence Act) 2007 (NSW) For the purposes of this Act, ‘intimidation’ of a person means conduct (including cyberbullying) amounting to harassment of the person. In this Act, ‘stalking’ includes contacting or otherwise approaching a person using the internet or any other technologically assisted means. A person who stalks or intimidates another person with the intention of causing the other person to fear physical or mental harm is guilty of an offence. Imprisonment for 5 years, 50 penalty units (equivalent to $5,500 as of June 2021) or both.

NT Stalking Section 189 of the Criminal Code Act 1983 (NT) A person (the offender) stalks another person (the victim) if the offender engages in conduct that includes repeated instances of or a combination of any of the following:
(a) telephoning, sending electronic messages to, or otherwise contacting, the victim or another person.
(b) giving offensive material to the victim or another person or leaving it where it will be found by, given to or brought to the attention of, the victim or the other person.
(c) acting in any other way that could reasonably be expected to arouse apprehension or fear in the victim for his or her own safety or that of another person. Imprisonment for 2 years.

SA Stalking Section 196AA of the Criminal Law Consolidation Act 1935 (SA) A person stalks another if on at least two separate occasions, the person:
(a) gives or sends offensive material to the other person or leaves offensive material where it will be found by, given to, or brought to the attention of the other person which can be transmitted by means of the internet, mail, telephone, facsimile or some other form of electronic communication.
(b) acts in any other way that could reasonably be expected to arouse the other person’s apprehension or fear; and
And the person:
(a) intends to cause serious physical or mental harm to the other person or a third person; or
(b) intends to cause serious apprehension or fear. Imprisonment for 3 years for a basic offence or imprisonment for 5 years for an aggravated offence.

TAS Stalking and bullying Section 192 of the Criminal Code Act 1924 (Tas) A person who, with intent to cause another person physical or mental harm, including self-harm, or extreme humiliation or to be apprehensive or fearful, pursues a course of conduct made up of one or more of the following actions: Imprisonment for 21 years.
(a) making threats to the other person or a third person
(b) directing abusive or offensive acts towards the other person or a third person
(c) sending offensive material to the other person or a third person or leaving offensive material where it is likely to be found by, given to, or brought to the attention of the other person or a third person
(d) publishing or transmitting offensive material by electronic or any other means in such a way that the offensive material is likely to be found by, or brought to the attention of, the other person or a third person
(e) using the internet or any other form of electronic communication in a way that could reasonably be expected to cause the other person to be apprehensive or fearful
(f) contacting the other person or a third person by postal, telephonic, electronic or any other means of communication
(g) acting in another way that could reasonably be expected to cause the other person physical or mental harm, including self-harm, or extreme humiliation or to be apprehensive or fearful is guilty of a crime.

| VIC | Stalking | Section 21A of the Crimes Act 1958 (Vic) | A person must not stalk another person. A person (the offender) stalks another person (the victim) if the offender engages in a course of conduct which includes any of the following:
(a) contacting the victim or any other person by post, telephone, fax, text message, e-mail, or other electronic communication or by any other means whatsoever.
(b) publishing on the Internet or by an e-mail or other electronic communication to any person a statement or other material
(ii) relating to the victim or any other person; or
(iii) purporting to relate to, or to originate from, the victim or any other person.
(c) tracing the victim's or any other person's use of the Internet or of e-mail or other electronic communications.
(d) making threats to the victim.
(e) using abusive or offensive words to or in the presence of the victim.
(f) using abusive or offensive words to or in the presence of the victim.
(g) giving offensive material to the victim or any other person or leaving it where it will be found by, given to or brought to the attention of, the victim or the other person.
(h) acting in any other way that could reasonably be expected
(i) to cause physical or mental harm to the victim, including self-harm; or
(ii) to arouse apprehension or fear in the victim for his or her own safety or that of any other person. | Imprisonment for 10 years. |

| WA | Stalking | Section 338E of the Criminal Code Act | There are two separate offences for stalking, with the first being: | Imprisonment for 1 year and a fine of $12,000. |
**Compilation Act 1913 (WA)**

A person who pursues another person with intent to intimidate that person, or a third person, is guilty of a crime.

<table>
<thead>
<tr>
<th>Country</th>
<th>Section/Act</th>
<th>Description</th>
<th>Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>WA</td>
<td>Section 338E of the Criminal Code Act Compilation Act 1913 (WA)</td>
<td>A person who pursues another person in a manner that could reasonably be expected to intimidate, and that does in fact intimidate that person or a third party, is guilty of a simple offence.</td>
<td>Imprisonment for 3 years.</td>
</tr>
</tbody>
</table>

### Unlawful Laws for Stalking

<table>
<thead>
<tr>
<th>State</th>
<th>Unlawful stalking</th>
<th>Description</th>
<th>Penalty</th>
</tr>
</thead>
</table>
| QLD   | Section 359B of the Criminal Code Act 1899 (Qld) | Unlawful stalking is conduct:  
(a) intentionally directed at a person (the "stalked person"); and  
(b) engaged in on any 1 occasion if the conduct is protracted or on more than 1 occasion, and  
(c) consisting of 1 or more acts of the following, or a similar, type  
(i) contacting a person in any way, including, for example, by telephone, mail, fax, email or through the use of any technology  
(ii) leaving offensive material where it will be found by, given to or brought to the attention of, a person  
(iii) giving offensive material to a person, directly or indirectly  
(iv) an intimidating, harassing or threatening act against a person, whether or not involving violence or a threat of violence  
(v) an act of violence, or a threat of violence, against, or against property of, anyone, including the defendant; and  
(d) that (i) would cause the stalked person apprehension or fear, reasonably arising in all the circumstances, of violence to, or against property of, the stalked person or another person; or (ii) causes detriment, reasonably arising in all the circumstances, to the stalked person or another person. | Unlawful act. No criminal charge applicable. |

There are no specific federal laws for stalking. Queensland considered stalking as "unlawful" and it is therefore covered under civil laws.

With the Tasmanian 21-year maximum penalty removed as an outlier, Victoria had the highest maximum penalty for stalking of 10 years imprisonment. One jurisdiction had five years, two had three years, two had two years and one had one year imprisonment respectively. With an average maximum penalty of 3.71 years imprisonment, the data suggests the offense can attract a substantial punishment term and may indicate that Australian law considers stalking a ‘moderate’ crime. However, stalking is considered unlawful (not illegal) in Queensland and therefore no criminal charge or jail time is applicable. For this reason, Queensland was not included in the quantitative analysis above. The broad range of maximum penalties between ten years and no criminal charge may imply
a national indecisiveness about the seriousness of the behaviour that may need to be addressed with a more universal approach.

The internet is a public place in which laws that govern predatory behaviour in real-life can be applied however, the internet can be written and read from anywhere and therefore does not conform to jurisdictional boundaries. No criminal sanction is attached for stalking under section 359B of the Criminal Code Act 1899 (Qld), obstructing any grounds for police involvement despite how severe the intimidating conduct may be. Remedies without potential imprisonment may not adequately address the impact of malicious attacks on victims or appropriately support journalists in a workplace often intertwined with online abuse. A universal approach to stalking laws may more effectively deter abusers.

Stalking has the potential to seep from online restraints to real-life scenarios that jeopardise the personal safety of journalists. Escalation of such behaviours can be demonstrated in the example of convicted Australian troll Zane Alchin, whose slut-shaming of a girl on Tinder in 2015 evolved into breaking into the woman's home before demanding an apology in her son's blood in 2020 (Sutton, 2020).

On 18 February 2021, an online petition by Sydney schoolgirl Chanel Contos demanded change in Australia's sex education curriculum (Hughes, 2021). Thousands of current and former students shared stories of the sexual violence they experienced as school students, and the petition received over 43,000 signatures (Teach Us Consent, 2021), which demonstrated this was a serious problem at all levels of Australian society (Hughes, 2021). The petition prompted schools to provide better education on consent earlier, from year three to year ten (Hendriks, 2010). While this was not a change in legislation, control of the school curriculum sits in the governance level of the Trolling Paradigm and so it is appropriate to mention here. Better and more actively enforced laws for online harassment that current policing does not prioritise may have the potential to reduce more serious future crimes through early intervention strategies. Future research may identify these online threats as precursors to help predict incidents.
that are likely to escalate to real-world violence and equip journalists with the necessary tools to combat the developing threat (Posetti, Harrison & Waisbord, 2020).

**Criminal laws for obscenity**

Encompassed by the criminal law theme of obscenity, five sub-categories were identified that include offensive language, offensive behaviour, disorderly conduct, indecent material, and indecent exposure. None are embedded in federal offences. An example of trolling in breach of each respective obscenity law is also provided in the descriptive tables of Figures 43, 44 and 45 in this section. International approaches to obscenity laws differ in severity; Germany has taken a more resolute stance against both trolls and the platforms that host them (Kirschbaum, 2016). The German legislation imposes heavy financial penalties on social media companies whose self-regulation fails to delete abusive material within 24 hours (McGoodan, 2017). The strict law requires hosting sites to take down hate-speech and other forms of illegal speech within twenty-four hours or face a potential 50-million-euro fine (McGoodan, 2017). Although critics suggest tough laws lead to “overbroad censorship” and “privatises enforcement of the law” by putting it in the hands of social media companies (Gorman, 2019, p. 259), ultimately there is a fine balancing act when it comes to legislation. Complex and confronting questions about the appropriate considered of free speech must be discussed, scrutinised, and addressed.

**Figure 43**

*Australian criminal laws for offensive language that trolled journalists could use derived from chapter 4.7 data*

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
<th>Maximum Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSW</td>
<td>Offensive language</td>
<td>Section 4A of the Summary Offences Act 1988 (NSW)</td>
<td>A person must not use offensive language in or near, or within hearing from, a public place or a school.</td>
<td>6 penalty units (equivalent to $660 as of June 2021)</td>
</tr>
<tr>
<td>NT</td>
<td>Obscenity</td>
<td>Section 53 of the Summary Offences Act 1923 (NT)</td>
<td>Any person who, in a public place, or within the view or hearing of any person passing therein: sings any obscene song or ballad, or writes or draws any indecent or obscene word, figure or representation, or uses any profane, indecent, or obscene language, shall be guilty of an offence.</td>
<td>Imprisonment for 6 months, a fine of $2,000, or both.</td>
</tr>
<tr>
<td>------</td>
<td>------------------------------</td>
<td>--------------------------------------------------</td>
<td>-------------------------------------------------------------------------------------------------</td>
<td>---------------------------------------------------</td>
</tr>
<tr>
<td>QLD</td>
<td>Offensive, obscene, indecent, or abusive language</td>
<td>Section 6 of the Summary Offences Act 2005 (Qld)</td>
<td>A person must not commit a public nuisance offence. A person commits a public nuisance offence if the person behaves in an offensive way. A person behaves in an offensive way if: (a) the person uses offensive, obscene, indecent, or abusive language; and (b) a person behaves in a threatening way if the person uses threatening language.</td>
<td>Imprisonment for 6 months or 10 penalty units (equivalent to $1,334.50 as of June 2021)</td>
</tr>
<tr>
<td>SA</td>
<td>Indecent language</td>
<td>Section 22 of the Summary Offences Act 1953 (SA)</td>
<td>A person who uses indecent or profane language or sings any indecent or profane song or ballad (a) in a public place; or (b) which is audible from a public place; or (c) with intent to offend or insult any person, is guilty of an offence.</td>
<td>Fine of $250.</td>
</tr>
<tr>
<td>TAS</td>
<td>Prohibited language</td>
<td>Section 12 of the Police Offenses Act 1935 (Tas)</td>
<td>A person shall not, in any public place, or within the hearing of any person in that place: (a) curse or swear (b) use any profane, indecent, obscene, offensive, or blasphemous language; or (c) use any threatening, abusive, or insulting words or behaviour calculated to provoke a breach of the peace or whereby a breach of the peace may be occasioned.</td>
<td>Imprisonment for 3 years or 3 penalty points (equivalent to $516 as of June 2021)</td>
</tr>
<tr>
<td>VIC</td>
<td>Obscene, indecent, threatening language</td>
<td>Section 17 of the Summary Offences Act 1966 (Vic)</td>
<td>Any person who in or near a public place or within the view or hearing of any person being or passing therein or thereon (a) writes or draws exhibits or displays an indecent or obscene word figure or representation; (b) uses profane indecent or obscene language or threatening abusive or insulting words; or (c) behaves in a riotous indecent offensive or insulting manner (including behaviour that involves a person exposing (to any extent) the person's anal or genital region) shall be guilty of an offence.</td>
<td>Imprisonment for 2 months or 10 penalty units (equivalent to $1,652.20 as of June 2021) for first offence. Imprisonment for 3 months or 15 penalty units (equivalent to $2,478.30 as of June 2021) for second offense. Imprisonment for 6 months or 25 penalty units (equivalent to $4,130.50 as of June 2021) for third offense.</td>
</tr>
</tbody>
</table>
The Australian Capital Territory and Western Australia do not have any specific laws in response to offensive language. Tasmania (included in this data set analysis) had the highest maximum penalty of 3 years imprisonment or three penalty units (equivalent to $516 as of June 2021) for offensive language. Four jurisdictions had the option of imprisonment for a period of 6 months or less, or financial reprimands. The remaining two jurisdictions had fines only. Although no quantitative analysis can conclude an average sentence, the lack of criminal defences for offensive language in two jurisdictions coupled with the potential for perpetrators to avoid jail time in all jurisdictions suggests use of offensive language is a ‘moderate’ crime. However, a broad range of maximum penalties, from 3 years imprisonment to no relevant applicable law, may imply a national indecisiveness about the severity of the behaviour or change social attitude to this type of language.

Figure 44

*Australian criminal laws for offensive behaviour that trolled journalists could use derived from chapter 4.7 data*

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
<th>Maximum Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>Offensive behaviour</td>
<td>Section 392 of the Crimes Act 1900 (ACT)</td>
<td>A person shall not in, near, or within the view or hearing of a person in, a public place behave in a riotous, indecent, offensive or insulting manner.</td>
<td>20 penalty units (equivalent to $3,200 as of June 2021)</td>
</tr>
<tr>
<td>NSW</td>
<td>Offensive conduct</td>
<td>Section 4 of the Summary Offences Act 1988 (NSW)</td>
<td>A person must not conduct himself or herself in an offensive manner in or near, or within view or hearing from, a public place or a school.</td>
<td>Imprisonment for 3 months or 6 penalty units (equivalent to $660 as of June 2021)</td>
</tr>
<tr>
<td>NT</td>
<td>Offensive conduct</td>
<td>Section 47 of the Summary</td>
<td>Every person who is guilty: (a) of any riotous, offensive, disorderly, or indecent behaviour, or of fighting, or using</td>
<td>Imprisonment for 6 months, a fine of</td>
</tr>
<tr>
<td>State</td>
<td>Offence</td>
<td>Act</td>
<td>Description</td>
<td>Penalty</td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
<td>-----</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>NT</td>
<td>Obscene language, in or within the hearing or view of any person in any road, street, thoroughfare or public place; (b) of unreasonably causing substantial annoyance to another person; or (c) of unreasonably disrupting the privacy of another person, shall be guilty of an offence.</td>
<td>Offences Act 1923 (NT)</td>
<td>$2,000, or both.</td>
<td></td>
</tr>
<tr>
<td>QLD</td>
<td>Offensive behaviour</td>
<td>Section 227 of the Criminal Code Act 1899 (Qld)</td>
<td>Any person who: (a) wilfully and without lawful excuse does any indecent act in any place to which the public are permitted to have access, whether on payment of a charge for admission or not; or (b) wilfully does any indecent act in any place with intent to insult or offend any person is guilty of a misdemeanour.</td>
<td>Imprisonment for 2 years.</td>
</tr>
<tr>
<td>SA</td>
<td>Indecent behaviour</td>
<td>Section 23 of the Summary Offences Act 1953 (SA)</td>
<td>A person who behaves in an indecent manner (a) in a public place, or while visible from a public place, or in a police station; or (b) in a place, other than a public place or police station, so as to offend or insult any person, is guilty of an offence.</td>
<td>Fine of $250.</td>
</tr>
<tr>
<td>TAS</td>
<td>Prohibited behaviour</td>
<td>Section 21 of the Police Offences Act 1935 (Tas)</td>
<td>A person must not, wilfully and without reasonable excuse, do any act or behave in a manner that a reasonable person is likely to find indecent or offensive in all the circumstances, if that person knew or should have known that his or her conduct was being, or may have been, viewed by another person.</td>
<td>Imprisonment for 1 year, 50 penalty points (equivalent to $8,600 as of June 2021), or both.</td>
</tr>
<tr>
<td>VIC</td>
<td>Obscene, indecent, threatening language and behaviour</td>
<td>Section 17 of the Summary Offences Act 1966 (Vic)</td>
<td>Any person who in or near a public place or within the view or hearing of any person being or passing therein or thereon (a) sings an obscene song or ballad; (b) writes or draws exhibits or displays an indecent or obscene word figure or representation; (c) uses profane indecent or obscene language or threatening abusive or insulting words; or (d) behaves in a riotous indecent offensive or insulting manner shall be guilty of an offence.</td>
<td>Imprisonment for 2 months or 10 penalty units (equivalent to $1,652.20 as of June 2021) for first offence. Imprisonment for 3 months or 15 penalty units (equivalent to $4,130.50 as of June 2021) for second offence. Imprisonment for 6 months or 25 penalty units (equivalent to $2,478.30 as of June 2021) for third or subsequent offence.</td>
</tr>
</tbody>
</table>
Offensive behaviour laws have the same fine (in South Australia) or higher penalties (in all other jurisdictions) as offensive language laws. Western Australia had the highest maximum penalty of three years imprisonment, Queensland had two years, with all other six jurisdictions enabling the avoidance of jail time with fines. Although no quantitative analysis can conclude an average sentence, data suggests the Australian law considers offensive behaviour a ‘moderate’ crime.

Figure 45

*Australian criminal laws for disorderly conduct that trolled journalists could use derived from chapter 4.7 data*

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
<th>Maximum Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>WA</td>
<td>Obscene acts in public</td>
<td>Section 202 of the Criminal Code Act Compilation Act 1913 (WA)</td>
<td>A person who does an obscene act in a public place or in the sight of any person who is in a public place.</td>
<td>Imprisonment for 3 years</td>
</tr>
<tr>
<td>NT</td>
<td>Disorderly behaviour</td>
<td>Section 47 of the Summary Offences Act 1923 (NT)</td>
<td>Every person who is guilty: (a) of any riotous, offensive, disorderly, or indecent behaviour, or of fighting, or using obscene language, in or within the hearing or view of any person in any road, street, thoroughfare or public place. (b) of unreasonably causing substantial annoyance to another person; or (c) of unreasonably disrupting the privacy of another person, shall be guilty of an offence.</td>
<td>Imprisonment for 6 months or fine of $2,000, or both.</td>
</tr>
<tr>
<td>QLD</td>
<td>Public nuisance</td>
<td>Section 6 in the Summary Offences Act 2005 (Qld)</td>
<td>A person must not commit a public nuisance offence. A person commits a public nuisance offence if: (a) the person behaves in a disorderly way; or (iv) an offensive way; or (v) a threatening way; or (vi) a violent way; and (b) the person’s behaviour interferes, or is likely to interfere, with the peaceful passage through, or enjoyment of, a public place by a member of the public.</td>
<td>Imprisonment for 6 months or 10 penalty points (equivalent to $1,334.50 as of June 2021)</td>
</tr>
<tr>
<td>SA</td>
<td>Disorderly behaviour</td>
<td>Section 7 of the Summary</td>
<td>A person who, in a public place or a police station:</td>
<td>Imprisonment for 3 months</td>
</tr>
</tbody>
</table>
Northern Territory and Queensland had the highest maximum penalty of six months imprisonment, South Australian and Tasmania had three months’ imprisonment, and Victoria and Western Australia had fines only. There were no specific laws in the Australian Capital Territory or New South Wales.

While a single abusive comment can be unpleasant or offensive, it may not cause a victim notable distress. However, a challenge of public spaces online is the potential onslaught of harassment a single person can be subjected to from multiple users over time (Jane, 2015). As such, legal responses can be difficult and often ineffective for journalists to adopt against multiple trolls (Powell & Henry, 2015). Although academic research has found that women and non-heterosexuals are more likely to experience abusive behaviours online (Powell & Henry, 2015), the potential impact of trolling extends to all journalists who navigate the online space as a fundamental part of their job. Extended periods of harassment can discourage victims and inhibit equal participation. However, as extensive parts of our lives are fused with online interactions, both socially and professionally, the
law should provide protection.

Thomas et al (2015) compared trolling on public websites with crimes involving bystander apathy. In the context of trolling, they defined active bystanders as key stakeholders who can challenge harassment and abuse through reporting violations of community standards and speaking up in support of victims. They described social media companies as ‘corporate bystanders’ that can curb disorderly conduct by enforcing policies. As explained in chapter 4.6 of this thesis (on p. 202), social media hosts can improve responses to trolling both proactively with the implementation of artificial intelligence, and reactively through quicker removal of offensive commentary after a complaint has been made. Powell and Henry (2015) encourage social media companies to take more proactive steps to create meaningful and enforceable community standards as part of a synthesised systems approach combating the malicious abuse they host.

**Criminal laws for indecency**

Indecency laws are presented in Figures 46 and 47 below. An example of trolling in breach of the law is also provided.

**Figure 46**

*Australian criminal laws for indecent material that trolled journalists could use derived from chapter 4.7 data*

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
<th>Maximum Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA</td>
<td>Indecent or offensive material</td>
<td>Section 33 of the Summary Offences Act 1953 (SA)</td>
<td>A person who (a) deposits indecent or offensive material in a public place or, except with the permission of the occupier, in or on private premises; or (b) exhibits indecent material to a person so as to offend or insult that person; or (c) delivers or exhibits indecent or offensive material to a minor (other than a minor of whom the person is a parent or guardian) is guilty of an offence.</td>
<td>Imprisonment for 6 months or fine of $20,000.</td>
</tr>
</tbody>
</table>

There are no specific federal laws or state laws in ACT, NSW, NT, QLD, TAS, VIC and WA for indecent material.

*Chapter 4.7 Legal responses results and discussion*
South Australia is the only jurisdiction to specify a standalone law for indecent material; however, trolling content of this nature has the potential to fall under the three umbrella laws governing obscenity for other jurisdictions.

**Figure 47**

*Australian criminal laws for indecent exposure that trolled journalists could use derived from chapter 4.7 data*

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
<th>Maximum Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>Indecent exposure</td>
<td>Section 392 of the Crimes Act 1900 (ACT)</td>
<td>A person who offends against decency by the exposure of his or her person in a public place, or in any place within the view of a person who is in a public place, commits an offence.</td>
<td>Imprisonment for 1 year, 20 penalty units (equivalent to $3,200 as of June 2021), or both.</td>
</tr>
<tr>
<td>NSW</td>
<td>Obscene exposure</td>
<td>Section 5 of the Summary Offences Act 1988 (NSW)</td>
<td>A person shall not, in or within view from a public place or a school, wilfully and obscenely expose his or her person.</td>
<td>Imprisonment for 6 months or 10 penalty units (equivalent to $1,100 as of June 2021)</td>
</tr>
<tr>
<td>NT</td>
<td>Indecent exposure</td>
<td>Section 50 of the Summary Offences Act 1923 (NT)</td>
<td>Any person who offends against decency by the exposure of his person in any street or public place, or in the view thereof, shall be guilty of an offence.</td>
<td>Imprisonment for 6 months, a fine of $2,000, or both.</td>
</tr>
<tr>
<td>QLD</td>
<td>Wilful exposure</td>
<td>Section 9 in the Summary Offences Act 2005 (Qld)</td>
<td>A person in a public place must not wilfully expose his or her genitals, unless the person has a reasonable excuse. It is a circumstance of aggravation for this section for a person to wilfully expose his or her genitals so as to offend or embarrass another person.</td>
<td>Imprisonment for 1 year or 40 penalty points (equivalent to $5,338 as of June 2021)</td>
</tr>
<tr>
<td>SA</td>
<td>Indecent behaviour</td>
<td>Section 23 of the Summary Offences Act 1953 (SA)</td>
<td>A person who, in a public place, or while visible from a public place or from occupied premises, wilfully does a grossly indecent act, whether alone or with another person, is guilty of an offence.</td>
<td>Imprisonment for 6 months or a fine of $2,500.</td>
</tr>
<tr>
<td>TAS</td>
<td>Prohibited behaviour</td>
<td>Section 21 of the Police Offences Act 1935 (Tas).</td>
<td>A person must not, wilfully and without reasonable excuse, do any act or behave in a manner that a reasonable person is likely to find indecent or offensive in all the circumstances, if that person knew or should have known that his or her conduct was</td>
<td>Imprisonment for 1 year, 50 penalty points (equivalent to $8,600 as of</td>
</tr>
</tbody>
</table>
Chapter 4.7 Legal responses results and discussion

Victoria had the highest maximum penalty for indecent exposure of two years imprisonment, four jurisdictions had one year imprisonment or a fine, and three had six months imprisonment or a fine. With an average maximum penalty of 0.93 years imprisonment, the data suggests the offense can attract a lower punishment term and may indicate that Australian law considers indecent exposure a ‘moderate’ crime.

In response to escalating incidents of indecent exposure online, the Australian government introduced Federal intimate image laws in 2018 (De Fina & Haider, 2019. The Enhancing Online Safety Act (Non-Consensual Sharing of Intimate Images) Act 2018 (Cth) amends the Criminal Code Act 1995 (Cth) by inserting section 474.17A. Also known as revenge porn laws, but more accurately known as image-based sexual abuse laws, the section criminalised the transmission, making available, publication, distribution, advertisement, or promotion of “private sexual material” as defined in the Act, without the affected person’s consent. Although their introduction received significant media traction, their relevance is not widely applicable to journalists whose abusers are more often strangers or former associates who would not have access to naked photos of them. For this reason, these laws were excluded from the data set.

A study by the International Women’s Media Foundation (2018) stated “overwhelming” (Ferrier, 2018, p. 48) female journalists report regularly receiving sexually explicit images, comments, emails, or text messages during the
course of a regular day at work, as reflected in the findings in chapter 4.1 of this study (on p. 30). Ranging from unsolicited photographs of the sender’s genitals to repeated sexual requests, the constant barrage of graphic content has become a normalised “common part” of contemporary journalism (West, 2015, para. 2). Anonymity is a reoccurring obstruction to the prosecution of perpetrators. In these cases, legislative action against social media organisations is the only available legal avenue. This strategy has had limited success in the Australian courts.

Critics have suggested that the burden of prosecution for trolling should not be placed exclusively on victims, and law enforcement officers should play a more active role in policing applicable laws (Citron, 2014; Greig, 2016; Jane 2015; Gorman, 2019; Molan 2020). Because they are so rarely used successfully, indecency laws have been described as “pathetic”, “piecemeal” and “completely inadequate” by Nigel Phair, the director of Centre for Internet Safety at the University of Canberra and former long-time Australian Federal Police officer (Phair, as cited in Gorman, 2019, p. 115). As a person with extensive knowledge and personal experience on such an issue, Phair expressed grave concerns about current law enforcement processes. When it comes to police, he suggested, “their days are full of policing terrestrial crimes” and frankly they are “still trying to come to grapple with what cyberhate is... let alone building the capacity and capability to investigate such matters” (p. 115). Australian police may require further training about application of relevant laws to trolling incidents as suggested in interviews with Mark Pearson and Mark Polden and other supporting literature (Jane 2015; Greig, 2016; Gorman, 2019; Molan 2020).

**Criminal laws for vilification**

There are two types of vilification: unlawful vilification, which is a civil matter, and serious vilification, which is a criminal offence. Unlawful vilification holds no criminal charge and is covered under Australian civil laws. If the unlawful vilification includes a threat of harm to a person or their property, or inciting others to threaten physical harm to a person or their property, it is a criminal offence. This is called serious vilification and is a police matter (Queensland
Human Rights Commission, 2021). For ease of comparison, the study grouped all vilification laws together within the civil law discussion section.

**Flowchart of Australian civil laws**

Two critical requirements must be met before a civil legal pathway can commence. These requirements are: the identity of the troll must be known, and troll must be located in Australia. An additional key consideration is the economic status of the troll. Civil law, also known as private law, concerns the rights and obligations of journalists in their relationship with other individuals and companies. As explained by Pearson and Polden (2015): “such cases usually result in courts ordering the unsuccessful party to perform or refrain from some action or make some payment in compensation for damage they have caused” (p. 66). These actions can include publishing apologies or retractions, refraining from publishing or removing online content. If the abuser does not have adequate finances to warrant pursuing damages, a lawsuit against the social media company that hosts the trolling content could be considered. If removal of the content is a priority, a civil case may still be pursued at the expense of the victim, the employer or insurance. Debate surrounds the fairness of the current Australian civil laws, which can burden the victim with hefty legal fees (Gittins, 2017), further straining the already delicate state of their mental health. A possible solution to this would be payment of legal fees by media employers, but that would still leave freelance journalists vulnerable. If a troll is being sued, breaching, defying, or ignoring a civil court order may become a criminal matter, by way of the law of contempt (Pearson & Polden, 2019). These three key considerations were included in the constructed flowchart of Australian civil laws displayed in Figure 48 over the two pages that follow.
Figure 48 Flowchart of civil laws that trolled journalists could use derived from chapter 4.7 data

Is the identity of the troll known?

- YES
- NO

Is the troll in Australia?

- YES
- NO

Do you wish to proceed with criminal charges?

- YES
- NO

Does the troll have a sufficient economic status to warrant pursuing damages?

- YES
- NO

What type of civil trolling was received?

Potential civil case against the social media company for defamation

Defamation

Personal injury

Claims against troll

Damages for non-economic loss (Pain and suffering)

FEDERAL LAWS (General anti-discrimination laws that do not specify civil claims)
Australian Human Rights Commission Act 1986 (Cth)
Disability Discrimination Act 1992 (Cth)
Discrimination Act 2004 (Cth)
Racial Discrimination Act 1975 (Cth)
Sex Discrimination Act 1984 (Cth)

STATE LAWS
ACT - Section 99 of the Civil Law (Wrongs) Act 2002 (ACT)
NSW - Section 16 of the Civil Liability Act 2002 (NSW)
NT - Section 17 of the Personal Injuries (Liabilities and Damages) Act 2003 (NT)
QLD - Section 62 of the Civil Liability Act 2003 (Qld)
SA - Section 54 of the Civil Liability Act 1996 (SA)
TAS - Section 28 of the Civil Liability Act 2002 (Tas)
VIC - Section 240 of the Wrongs Act 1958 (Vic)
WA - Section 9 of the Civil Liability Act 2002 (WA)

Damages for past economic loss

FEDERAL LAWS (General anti-discrimination laws that do not specify civil claims)
Australian Human Rights Commission Act 1986 (Cth)
Disability Discrimination Act 1992 (Cth)
Discrimination Act 2004 (Cth)
Racial Discrimination Act 1975 (Cth)
Sex Discrimination Act 1984 (Cth)

STATE LAWS
ACT - Section 33 of the Civil Law (Wrongs) Act 2002 (ACT)
NSW - Section 12 of the Civil Liability Act 2002 (NSW)
NT - Section 20 of the Personal Injuries (Liabilities and Damages) Act 2003 (NT)
QLD - Section 54 of the Civil Liability Act 2003 (Qld)
SA - Section 54 of the Civil Liability Act 1996 (SA)
TAS - Section 28 of the Civil Liability Act 2002 (Tas)
VIC - Section 241 of the Wrongs Act 1958 (Vic)
WA - Section 11 of the Civil Liability Act 2002 (WA)

Damages for future economic loss

FEDERAL LAWS (General anti-discrimination laws that do not specify civil claims)
Australian Human Rights Commission Act 1986 (Cth)
Disability Discrimination Act 1992 (Cth)
Discrimination Act 2004 (Cth)
Racial Discrimination Act 1975 (Cth)
Sex Discrimination Act 1984 (Cth)

STATE LAWS
ACT - Section 13 of the Civil Law (Wrongs) Act 2002 (ACT)
NSW - Section 13 of the Civil Liability Act 2002 (NSW)
NT - Section 20 and 21 of the Personal Injuries (Liabilities and Damages) Act 2003 (NT)
QLD - Section 57 of the Civil Liability Act 2003 (Qld)
SA - Section 55 of the Civil Liability Act 1996 (SA)
TAS - Section 28A of the Civil Liability Act 2002 (Tas)
VIC - Section 242 of the Wrongs Act 1958 (Vic)
WA - Section 11 of the Civil Liability Act 2002 (WA)

Damages for superannuation loss

FEDERAL LAWS (General anti-discrimination laws that do not specify civil claims)
Australian Human Rights Commission Act 1986 (Cth)
Disability Discrimination Act 1992 (Cth)
Discrimination Act 2004 (Cth)
Racial Discrimination Act 1975 (Cth)
Sex Discrimination Act 1984 (Cth)

STATE LAWS
NSW - Section 15C of the Civil Liability Act 2002 (NSW)
QLD - Section 16 of the Civil Liability Act 2003 (Qld)
TAS - Section 25 of the Civil Liability Act 2002 (Tas)
There are no specific laws in ACT, NT, SA, VIC and WA.
Descriptive tables of Australian civil laws

Through thematic analysis of qualitative and quantitative data results, the study identified three civil law themes of defamation, personal injury and vilification in addition to the seven criminal law themes described prior. Each theme was investigated to determine which Australian laws journalists could use to seek damages against a perpetrator under each individual category. No maximum penalty is listed as the amount of damages varies depending on the circumstances of each particular case. The study provides one example of trolling for each civil law theme. The findings are categorised by theme, then by jurisdiction, and are presented in the descriptive tables that follow. Each table is followed by a brief discussion to provide context for critical analysis.

Australian vilification laws

For ease of comparison, the study grouped all vilification laws together and briefly re-described their differences before the analysis and discussion in the following section. There are two types of vilification: unlawful vilification, which is a civil matter, and serious vilification, which is a criminal offence. Unlawful vilification that holds no criminal charge and is covered under Australian civil laws. If the unlawful vilification includes a threat of harm to a person or their property, or inciting others to threaten physical harm to a person or their property, it is a criminal offence. This is called serious vilification and is a police matter (Queensland Human Rights Commission, 2021). Australian laws divide vilification into six sub-categories that include racial, religious, sexuality, gender-identity, HIV/AIDS, and disability vilification. Both serious vilification (criminal) and civil vilification laws that journalists could use to prosecute a troll for each category will be discussed collectively.

Australian laws for racial vilification

As defined by the Australian Human Rights Commission (2021a) “racial discrimination is when a person is treated less favourably than another person in a similar situation because of their race, colour, descent, national or ethnic origin or immigrant status” (para. 1), whereas racial vilification is “doing something in public based on the race, colour, national or ethnic origin of a person or group of
people which is likely to offend, insult, humiliate or intimidate” (para. 6). Racial hatred is universally addressed as an offensive and prohibited act under Australian legislation, however, while such behaviour may be unlawful it is not necessarily a criminal offence under the legislation as described in in Figure 49 below. An example of trolling in breach of the law also provided.

**Figure 49**

*Australian laws for racial vilification that trolled journalists could use derived from chapter 4.7 data*

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
<th>Maximum Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>Vilification</td>
<td>Section 67A of the Criminal Code 2002 (ACT)</td>
<td>A person commits an offence if (a) the person intentionally carries out an act; and (b) the act is a threatening act; and (c) the person is reckless about whether the act incites hatred toward, revulsion of, serious contempt for, or severe ridicule of, a person or group of people on the ground of any of the following: (i) race (d) the act is done other than in private; and (e) the person is reckless about whether the act is done other than in private.</td>
<td>50 penalty points (equivalent to $8,000 as of June 2021).</td>
</tr>
<tr>
<td>SA</td>
<td>Racial vilification</td>
<td>Section 4 of the Racial Vilification Act 1996 (SA)</td>
<td>A person must not, by a public act, incite hatred toward, serious contempt for, or severe ridicule of, a person or group of persons on the ground of their race by (a) threatening physical harm to the person, or members of the group, or to property of the person or members of the group; or (b) inciting others to threaten physical harm to the person, or members of the group, or to property of the person or members of the group.</td>
<td>Imprisonment for 3 years, a fine of $5,000, or both.</td>
</tr>
<tr>
<td>WA</td>
<td>Conduct intended to incite racial animosity or racist harassment</td>
<td>Section 77 of the Criminal Code Act Compilation Act 1913 (WA)</td>
<td>Any person who engages in any conduct, otherwise than in private, by which the person intends to create, promote or increase animosity towards, or harassment of, a racial group, or a person as a member of a racial group, is guilty of a crime.</td>
<td>Imprisonment for 14 years.</td>
</tr>
<tr>
<td>WA</td>
<td>Conduct intended to racially harass</td>
<td>Section 80A of the Criminal Code Act</td>
<td>Any person who engages in any conduct, otherwise than in private, by which the person intends to harass a racial group, or a person as a member of a racial group, is guilty of a crime.</td>
<td>Imprisonment for 5 years.</td>
</tr>
</tbody>
</table>
### “Unlawful” Civil Laws

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Vilification Theme</th>
<th>Relevant Section</th>
<th>Legal Basis</th>
<th>Unlawful Act?</th>
<th>Criminal Charge Applicable?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federal</td>
<td>Unlawful offensive behaviour because of race, colour or national or ethnic origin</td>
<td>Section 18C of the Racial Discrimination Act 1975 (Cth)</td>
<td>It is unlawful for a person to do an act, otherwise than in private, if: (a) the act is reasonably likely, in all the circumstances, to offend, insult, humiliate or intimidate another person or a group of people; and (b) the act is done because of the race, colour or national or ethnic origin of the other person or of some or all of the people in the group.</td>
<td>Unlawful act.</td>
<td>No criminal charge applicable.</td>
</tr>
<tr>
<td>NSW</td>
<td>Unlawful racial vilification</td>
<td>Section 20C of the Anti-Discrimination Act 1977 (NSW)</td>
<td>It is unlawful for a person, by a public act, to incite hatred towards, serious contempt for, or severe ridicule of, a person or group of persons on the ground of the race of the person or members of the group.</td>
<td>Unlawful act.</td>
<td>No criminal charge applicable.</td>
</tr>
<tr>
<td>QLD</td>
<td>Unlawful racial vilification</td>
<td>Section 124A of the Anti-Discrimination Act 1991 (Qld)</td>
<td>A person must not, by a public act, incite hatred towards, serious contempt for, or severe ridicule of, a person or group of persons on the ground of the race, religion, sexuality or gender identity of the person or members of the group.</td>
<td>Unlawful act.</td>
<td>No criminal charge applicable.</td>
</tr>
<tr>
<td>SA</td>
<td>Racial victimisation</td>
<td>Section 73 of the Civil Liability Act 1936 (SA)</td>
<td>The act of racial victimisation means a public act inciting hatred, serious contempt or severe ridicule of a person or group of persons on the ground of their race but does not include (a) publication of a fair report of the act of another person; or (b) publication of material in circumstances in which the publication would be subject to a defence of absolute privilege in proceedings for defamation; or (c) a reasonable act, done in good faith, for academic, artistic, scientific or research purposes or for other purposes in the public interest (including reasonable public discussion, debate or expositions)</td>
<td>The total amount of the damages that may be awarded for the same act or series of acts cannot exceed $40,000.</td>
<td></td>
</tr>
<tr>
<td>TAS</td>
<td>Unlawful inciting hatred</td>
<td>Section 19 of the Anti-Discrimination Act 1998 (Tas)</td>
<td>A person, by a public act, must not incite hatred towards, serious contempt for, or severe ridicule of, a person or a group of persons on the ground of the race of the person or any member of the group.</td>
<td>Unlawful act.</td>
<td>No criminal charge applicable.</td>
</tr>
<tr>
<td>VIC</td>
<td>Unlawful racial vilification</td>
<td>Section 7 of the Racial and Religious Tolerance Act 2001 (Vic)</td>
<td>A person must not, on the ground of the race of another person or class of persons, engage in conduct that incites hatred against, serious contempt for, or revulsion or severe ridicule of, that other person or class of persons.</td>
<td>Unlawful act.</td>
<td>No criminal charge applicable.</td>
</tr>
</tbody>
</table>

No specific religious vilification laws in relation to online abuse in NT.

Of the six vilification themes identified, racial vilification accrued the highest...
number of specified criminal laws, totalling four. Racial vilification was considered illegal in three jurisdictions. Western Australia had a significantly higher maximum penalty for inciting racial vilification of 14 years imprisonment. South Australia was the only other jurisdiction with potential jail time, set at three years. Five jurisdictions considered racial vilification an “unlawful act” to which no criminal charge is applicable. This may indicate that Australian law considers racial vilification a ‘less serious’ crime. On analysis of Australian Capital Territory law, the study identified section 72 of the Discrimination Act 1991 (ACT) clearly states, “unlawful act provides no basis for civil action”. This statement is noted as an outlier; unlawful acts do provide the basis for civil action under Federal, New South Wales, Queensland, Tasmania, and Victoria jurisdictions.

**Racial vilification laws are not often successful**

Each of the nine Australian jurisdictions addresses hate speech in different ways with varying levels of intensity, and the absence of a unified criminal policy highlights the fine line between opinion and bigotry. Although these laws are aimed to prohibit inciting hatred or ridicule, data from a 2021 report by the Victorian Equal Opportunity and Human Rights Commission revealed that in 20 years, only one successful racial vilification prosecutions had been successful in Victoria (Cotrell v Smith, 2019). A 2019 *ABC* article by Sarah Jane Bell, retrieved within the archival evidence of this chapter, was of particularly significant insight. Victorian Equal Opportunity and Human Rights Commissioner Kristen Hilton explained that current laws burdened the victim with the onus to prove the troll’s behaviour and would instigate further hateful conduct by a third person and this “threshold for meeting the vilification test is simply too high” (Hilton, as cited in Bell, 2021, para. 16) and “needed to be changed” (para. 19). Hilton explained “the reason many enquiries were not progressed was because, under the current act, the victim needed to be able to name the person who allegedly vilified the victim” (para. 21), which was often difficult to obtain for abuse that occurred online. Hilton labelled the current Victorian legislation “convoluted and cumbersome” (para. 31). Although public awareness of the impact of vilification is growing, online incidents continue to rise, indicating legal reform is required to find a better balance between free speech and vilification prevention: but legislation is only
part of the problem. As Ackoff’s (1974) System Theory that underpins this study suggests, legislative change in conjunction with community awareness and education is the key to more effective prevention strategies.

**Extensive exemption list**

Australian legislation has an extensive list of public interest exemptions to racial vilification laws. In a further amendment of the *Racial Discrimination Act 1975 (Cth)*, the *Racial Hatred Act 1995 (Cth)* permits multiple exclusions including words said in reasonably good faith in the performance or exhibition of artistic work; statements, publication and discussions of scientific and academic purpose or genuine public interest; and fair and accurate reporting. However, section 18D noted a further exemption: “a fair comment on any event or matter of public interest if the comment is an expression of a genuine belief held by the person making the comment.” The study questions the use of this exemption as a trolling defence and suggests abusive racial comments should be considered unlawful, irrelevant of the troll’s belief that the comments are true. A troll’s genuine belief of the truth of an abusive comment should not affect its unlawful status. Considering the trolling example of Australian AFL player Adam Goodes (Blackwell, 2015) who was frequently abused by opponents’ fans both verbally and online, the study suggests that whether Goodes’s abusers held a genuine belief that his appearance was similar to an ape should be irrelevant to their label as unlawful. The study suggests a more appropriate measure to determine the lawful status of abuse should be consideration of the pain and suffering these comments caused the victim. Although the study grants the notion the public should have a right to freely express their thoughts and opinions, this should not be permitted if it is at the expense of another person’s emotional and mental well-being. The study suggests the growing acknowledgement of the grave impacts racial abuse can have on victims may warrant a review of these laws.

**Outlier**

Racial hatred is universally addressed as an offensive and prohibited act under Australian legislation; however, while such behaviour may be unlawful, it is not necessarily a criminal offence under the legislation and does not necessarily
provide the grounds for civil action. Section 72 of the Discrimination Act 1991 (ACT) clearly states, “unlawful act provides no basis for civil action” and highlights such laws may more accurately be a list of guidelines, with minimal consequences for those who choose not to abide by them.

Andrew Bolt case study

Although racial vilification is ‘unlawful’ in Victoria, a notable Australian case study is the class action against journalist Andrew Bolt. Under section 18C of the Australian Human Rights Commission, a complaint was made about two articles by Andrew Bolt that implied light-skinned people who identified as Aboriginal did so for personal gain (Ritchie, 2011). The first ‘It’s so hip to be black’ was published by the Herald Sun newspaper and republished on websites The Herald and Weekly Times under the title ‘White is the new black’ on April 15, 2009 (Eatock v Bolt). The second article “White fellas in the black’ was published by the same news organisations on August 21, 2009, in both print and online (Eatock v Bolt). The Federal Court of Australia found that the article had breached section 18C and ordered that the Herald Sun and Weekly Times publish an article of equal prominence to publicise the court’s decision. The plaintiff, Pat Eatock, did not seek compensation and only sought a declaration from the court that the writing and publishing of the articles was unlawful, an apology, an order preventing republication, and for the articles to be taken down from websites. Therefore, Bolt was not prosecuted, convicted, fined, jailed, or even made to pay compensation for his actions. The news organisation (not Bolt) was ordered to pay substantial legal costs, although the exact amount was not made public.

The price of free speech is that society accepts that people should generally be able to make offensive commentary; however, there are limits. Education is a key strategy in building a culture of tolerance and non-discrimination, and the law is one useful tool that helps to achieve this by setting out appropriate standards of conduct. People are more likely to speak out in public against racism if the law supports their position and, conversely, are less likely to engage in racial vilification if the law makes it unlawful (Human Rights Law Centre, 2014). In this way, the law is an important tool that complements education to combat racism.
It is also important to provide access to legal remedies for victims of racial vilification. Groups that experience racial vilification are often unable to participate in the public debate on an equal footing with others and racial vilification can have the perverse impact of causing affected people and groups to retreat from public participation (Human Rights Law Centre, 2014). The Andrew Bolt case involved Australia’s most widely read columnist unreasonably and in bad faith engaging in conduct reasonably likely to racially offend, insult, humiliate and intimidate light-skinned Aboriginal people. The law provided an important tool to address this in a way that public debate was unable to.

**Australian laws for religious vilification**

Religious vilification is less frequently discussed than racial vilification but it is succinctly defined by Victorian Equal Opportunity and Human Rights Commission (2021) as: “when a person behaves in a way that encourages hatred, revulsion or ridicule of another person, because of the other person’s religion” (para. 3). Such behaviours can diminish the dignity and self-worth of a victim, affect an individual’s ability to contribute to society, and undermine community harmony and social cohesion. Australian religious vilification laws are presented in Figure 50 below with an example of trolling in breach of the law also provided.

**Figure 50**

*Australian laws for religious vilification that trolled journalists could use derived from chapter 4.7 data*

<table>
<thead>
<tr>
<th>Australian Laws for Religious Vilification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Trolling Example:</strong> “Pooper Muslims like you receive the death penalty in your country, go home”</td>
</tr>
<tr>
<td>(Participant 5 in section 4.2 of this study)</td>
</tr>
</tbody>
</table>

Section 46P of the Human Rights and Equal Opportunity Commission Act 1986 (Cth) allows journalists to make complaints to the Race Discrimination Commissioner Chin Tan (in 2021) or Aboriginal and Torres Strait Islander Social Justice Commissioner June Oscar (in 2021).

<table>
<thead>
<tr>
<th>Criminal Laws</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Jurisdiction</strong></td>
</tr>
<tr>
<td>ACT</td>
</tr>
</tbody>
</table>
At a federal level, Section 18C of the *Racial Discrimination Act 1975 (Cth)* makes it unlawful to offend someone based on their religious conviction, but not a crime. Serious religious vilification was considered illegal in the Australian Capital Territory only, with the maximum penalty consisting of a fine with no jail time. Religious vilification was considered an unlawful act to which no criminal charge was applicable and commencing civil claims for compensation is the only option in four jurisdictions. An additional four jurisdictions did not specify any law in relation to religious vilification. The data may indicate that Australian law considers religious vilification a ‘less serious’ crime.

Australian religious vilification laws are lenient in comparison to countries where religious expression is practised more widely as a fundamental part of social

---

**Chapter 4.7 Legal responses results and discussion**
culture. Focussing on Australia’s blasphemy laws, senior lecturer in constitutional law at Western Sydney University Luke Beck (2017) explains, “the crime of blasphemy is not about vilifying or inciting hatred against people on the basis of their religion, it is about protecting God and Christian doctrine from scurrilous commentary, and Christian religious sensibilities from offence” (para. 5). Therefore, Australia’s blasphemy laws, which Queensland and Western Australia abolished, are not included in this section, but are considered for comparison with more disciplinary international laws for global contextualisation. In 2021, eight Islamic countries of Nigeria, Pakistan, Iran, Afghanistan, Somalia, Mauritania, Saudi Arabia, and Brunei, it is a criminal offense to publicly doubt or deny the existence of God or Allah or to criticise religious laws such as Sharia. Offenses are considered capital offence (Humanists International, 2021). Although most of these countries have signed international agreements to treat all citizens equally, these laws fail to respect the rights of atheists and those with differing beliefs. Prison time can be served in an additional 47 countries for criticisms of faith, including academic study of the origins of religions. Although Australia is a non-denominational country, widespread multi-culturalism provides a rich and diverse range of religious expressions. Legislation should prioritise inclusiveness, respect cultural diversity, and protect the right to express ideologies free from abuse.

**Australian laws for disability vilification**

The definition of disability in the *Disability Discrimination Act 1992 (Cth)* is deliberately broad and does not require any assessment of the severity, type, permanency, or when or how it was acquired. In addition, the disability may not be current and could be in the past, future or imputed. Disability discrimination is a rare type of trolling about a journalist who is discriminated against for having a disability; however, it is hard to envision a circumstance in which the troll had the power to implement the discrimination. While disability vilification is also not considered commonly applicable, it was raised as a potential legal avenue in interviews with Mark Polden in this chapter and is therefore briefly explored. Of particular importance for this section, the study recognised spokesperson and Australian journalist Nas Campanella, who is the current disability affairs reporter
for the *ABC*. Australian disability vilification laws are presented in Figure 51 below.

**Figure 51**

*Australian laws for disability vilification that trolled journalists could use derived from chapter 4.7 data*

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
<th>Maximum Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>Vilification</td>
<td>Section 67A of the <em>Criminal Code 2002 (ACT)</em></td>
<td>A person commits an offence if (a) the person intentionally carries out an act; and (b) the act is a threatening act; and (c) the person is reckless about whether the act incites hatred toward, revulsion of, serious contempt for, or severe ridicule of, a person or group of people on the ground of any of the following: (i) disability (d) the act is done other than in private; and (e) the person is reckless about whether the act is done other than in private.</td>
<td>50 penalty points (equivalent to $8,000 as of June 2021)</td>
</tr>
</tbody>
</table>

"Unlawful" Civil Laws

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
<th>Maximum Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federal</td>
<td>Unlawful direct disability discrimination</td>
<td>Section 5 of the <em>Disability Discrimination Act 1992 (Cth)</em></td>
<td>For the purposes of this Act, a person (the discriminator) discriminates against another person (the aggrieved person) on the ground of a disability of the aggrieved person if, because of the disability, the discriminator treats, or proposes to treat, the aggrieved person less favourably than the discriminator would treat a person without the disability in circumstances that are not materially different.</td>
<td>Unlawful act. No criminal charge applicable.</td>
</tr>
<tr>
<td>TAS</td>
<td>Unlawful inciting hatred</td>
<td>Section 19 of the <em>Anti-Discrimination Act 1998 (Tas)</em></td>
<td>A person, by a public act, must not incite hatred towards, serious contempt for, or severe ridicule of, a person or a group of persons on the ground of any disability of the person or any member of the group.</td>
<td>Unlawful act. No criminal charge applicable.</td>
</tr>
</tbody>
</table>

No specific disability vilification laws in relation to online abuse NSW, NT, QLD, SA, VIC and WA

The study notes sustained abuse by a person or group would be required for a disability vilification claim to go to court. Disability vilification is illegal in the Australian Capital Territory only, with the maximum penalty consisting of a fine.
with no jail time. Disability vilification is stated as an unlawful act in two other jurisdictions (Federal and Tasmania), which are outlined under civil laws later in this chapter. Six jurisdictions did not specify any related law. The data may indicate that Australian law considers disability vilification a ‘less serious’ crime or that it is covered by other legislation.

The lack of universal disability anti-vilification laws in Australia highlights a gap in the national statutory regime and bridging this gap is suggested by Cullen, Mandikos and Vario (2010) as a starting point towards redressing society’s negative perceptions of people with disabilities, and thereby also addressing some of the great difficulties that confront people with disabilities. Although 1981 was formally recognised as the ‘International Year of Disabled Persons’ and was dedicated to raising public awareness of the systematic exclusion and stigmatisation of people with disabilities (Cullen, Mandikos & Vario, 2010), over 40 years later in 2021, the general attitude and social perception of people with disabilities remains largely unchanged. While derogatory comments such as ‘you are a spastic’ and ‘retarded’ (Chapter 4.2 findings on p. 54) are often made in jest, they are nevertheless evidence of a problem with the way disability is perceived within society, and remarks about disabilities are still used as an expression of contempt and derision in contemporary discourse (Tillers, 2020). Given that one in six, or 4.4 million Australians, live with a disability (Australian Institute of Health and Welfare, 2020) and 35.9% of Australia’s 8.9 million households include a person with disability, which includes mental health (Australian Network on Disability, 2021), the absence of protection by any criminal laws is disconcerting, considering this portion represents Australia’s largest minority group and arguably one of the most vulnerable.

**Australian laws for sexuality vilification**

Sexuality vilification is defined as: “communications made in public that incite hatred towards, serious contempt for, or severe ridicule of, a person or group of people on the ground of their sexual orientation” (Australian Human Rights Commission, 2021b) and may include graffiti, comments made on radio or television, web pages with public access or verbal abuse in a public place. These
laws are presented in Figures 52 below. Sexuality vilification, gender identity vilification and HIV/AIDS vilification are discussed in collaboration after the three tables that follow.

**Figure 52**

*Australian laws for sexual vilification that trolled journalists could use derived from chapter 4.7 data*

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
<th>Maximum Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>Vilification</td>
<td>Section 67A of the Criminal Code 2002 (ACT)</td>
<td>A person commits an offence if (a) the person intentionally carries out an act; and (b) the act is a threatening act; and (c) the person is reckless about whether the act incites hatred toward, revulsion of, serious contempt for, or severe ridicule of, a person or group of people on the ground of any of the following: (i) sex characteristics or sexuality (d) the act is done other than in private; and (e) the person is reckless about whether the act is done other than in private.</td>
<td>50 penalty points (equivalent to $8,000 as of June 2021)</td>
</tr>
</tbody>
</table>

**“Unlawful” Civil Laws**

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
<th>Unlawful act. No criminal charge applicable.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federal Unlawful discrimination on the ground of sexual orientation</td>
<td>Section 5A of the Sex Discrimination Act 1984 (Cth)</td>
<td>For the purposes of this Act, a person (the discriminator) discriminates against another person (the aggrieved person) on the ground of the aggrieved person’s sexual orientation if, by reason of: (a) the aggrieved person’s sexual orientation; or (b) a characteristic that appertains generally to persons who have the same sexual orientation as the aggrieved person; or (c) a characteristic that is generally imputed to persons who have the same sexual orientation as the aggrieved person; the discriminator treats the aggrieved person less favourably than, in circumstances that are the same or are not materially different, the discriminator treats or would treat a person who has a different sexual orientation.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSW Unlawful homosexual Vilification</td>
<td>Section 49ZT of the Anti-Discrimination</td>
<td>It is unlawful for a person, by a public act, to incite hatred towards, serious contempt for, or severe ridicule of, a person or group of persons</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Australian Laws for Sexuality Vilification*

Trolling Example: “Fat-arssed whore who can’t get laid” (Ford, 2016, p. 172)

Section 46P of the Human Rights and Equal Opportunity Commission Act 1986 (Cth) allows journalists to make complaints to the Sex Discrimination Commissioner, a position in 2021 held by Kate Jenkins.
<table>
<thead>
<tr>
<th>ACT</th>
<th>Legal response</th>
<th>Details</th>
<th>Charge applicable</th>
</tr>
</thead>
<tbody>
<tr>
<td>QLD</td>
<td>Unlawful sexuality vilification</td>
<td>Section 124A of the Anti-Discrimination Act 1991 (Qld) on the ground of the homosexuality of the person or members of the group.</td>
<td>Unlawful act. No criminal charge applicable.</td>
</tr>
<tr>
<td>TAS</td>
<td>Unlawful inciting hatred</td>
<td>Section 19 of the Anti-Discrimination Act 1998 (Tas) A person, by a public act, must not incite hatred towards, serious contempt for, or severe ridicule of, a person or group of persons on the ground of the sexual orientation or lawful sexual activity of the person or any member of the group.</td>
<td>Unlawful act. No criminal charge applicable.</td>
</tr>
</tbody>
</table>

No specific sexuality vilification laws in relation to online abuse in NT, SA, VIC and WA

Sexuality vilification is considered illegal in the Australian Capital Territory only, with the maximum penalty consisting of a fine with no jail time. Sexuality vilification is an unlawful act in four other jurisdictions (Federal, New South Wales, Queensland, and Tasmania). An additional four jurisdictions (Northern Territory, South Australia, Victoria, and Western Australia) did not specify any related law. The data may indicate that Australian law considers sexual vilification as a ‘less serious’ crime.

**Australian laws for gender identity vilification**

Gender identity vilification is defined as “communications made in public that incite hatred towards, serious contempt for, or severe ridicule of, a person or group of people on the ground of their sex and/or gender identity” (Australian Human Rights Commission, 2021b). Relevant Australian legislation is presented in Figure 53 below.
Figure 53

Australian laws for gender identity vilification that trolled journalists could use derived from chapter 4.7 data

<table>
<thead>
<tr>
<th>Australian Laws for Gender Identity Vilification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Trolling Example:</strong> “Trannies getting forgery won’t change your gender, it will only turn you into a deformed freak” (Gibbs, 2015)</td>
</tr>
</tbody>
</table>

Section 46P of the *Human Rights and Equal Opportunity Commission Act 1986 (Cth)* allows journalists to make complaints to the Sex Discrimination Commissioner, a position in 2021 held by Kate Jenkins.

<table>
<thead>
<tr>
<th><strong>Criminal Laws</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Jurisdiction</strong></td>
</tr>
<tr>
<td>ACT</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th><strong>“Unlawful” Civil Laws</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Jurisdiction</strong></td>
</tr>
</tbody>
</table>

<p>| <strong>Jurisdiction</strong> | <strong>Unlawful discrimination on the ground of intersex status</strong> | <strong>Section 5C of the Sex Discrimination Act 1984 (Cth)</strong> | For the purposes of this Act, a person (the discriminator) discriminates against another person (the aggrieved person) on the ground of the aggrieved person’s intersex status if, by reason of: (a) the aggrieved person’s intersex status; or (b) a characteristic that appertains generally to persons of intersex status; or (c) a characteristic that is generally imputed to persons of intersex status the discriminator treats the aggrieved person less favourably than, in circumstances that are the same or are not materially different, the discriminator treats or would treat a person who has a different gender identity. | Unlawful act. No criminal charge applicable. |</p>
<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Unlawful Vilification</th>
<th>Relevant Legislation</th>
<th>Description</th>
<th>Legal Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSW</td>
<td>Unlawful transgender vilification</td>
<td>Section 38S of the Anti-Discrimination Act 1977 (NSW)</td>
<td>It is unlawful for a person, by a public act, to incite hatred towards, serious contempt for, or severe ridicule of: (a) a person on the ground that the person is a transgender person, or (b) a group of persons on the ground that the members of the group are transgender persons.</td>
<td>Unlawful act. No criminal charge applicable.</td>
</tr>
<tr>
<td>QLD</td>
<td>Unlawful gender identity vilification</td>
<td>Section 124A of the Anti-Discrimination Act 1991 (Qld)</td>
<td>A person must not, by a public act, incite hatred towards, serious contempt for, or severe ridicule of, a person or group of persons on the ground of the race, religion, sexuality or gender identity of the person or members of the group.</td>
<td>Unlawful act. No criminal charge applicable.</td>
</tr>
<tr>
<td>TAS</td>
<td>Unlawful inciting hatred</td>
<td>Section 19 of the Anti-Discrimination Act 1998 (Tas)</td>
<td>A person, by a public act, must not incite hatred towards, serious contempt for, or severe ridicule of, a person or a group of persons on the ground of the gender identity or intersex variations of sex characteristics of the person or any member of the group.</td>
<td>Unlawful act. No criminal charge applicable.</td>
</tr>
</tbody>
</table>

No specific gender identity vilification laws in relation to online abuse in NT, SA, VIC and WA

Gender identity vilification is considered illegal in the Australian Capital Territory only, with the maximum penalty consisting of a fine with no jail time. Gender identity vilification is an unlawful act in four jurisdictions (Federal, New South Wales, Queensland, and Tasmania). An additional four jurisdictions (Northern Territory, South Australia, Victoria, and Western Australia) did not specify any related law. The data may indicate that Australian law considers gender identity vilification as a 'less serious' crime.

**Australian laws for HIV/AIDS vilification**

HIV/AIDS vilification is defined as “inciting hatred towards, revulsion of, serious contempt for, or severe ridicule of, a person or group of people on the ground of their HIV/AIDS status (meaning a person who has the Human Immunodeficiency Virus or Acquired Immune Deficiency Syndrome)” (Australian Human Rights Commission, 2021b) as presented in Figure 54 below.
Figure 54

Australian laws for HIV/AIDS that trolled journalists could use derived from chapter 4.7 data

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
<th>Maximum Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>Vilification</td>
<td>Section 67A of the Criminal Code 2002 (ACT)</td>
<td>A person commits an offence if (a) the person intentionally carries out an act; and (b) the act is a threatening act; and (c) the person is reckless about whether the act incites hatred toward, revulsion of, serious contempt for, or severe ridicule of, a person or group of people on the ground of any of the following: (i) HIV/AIDS status; (d) the act is done other than in private; and (e) the person is reckless about whether the act is done other than in private.</td>
<td>50 penalty points (equivalent to $8,000 as of June 2021)</td>
</tr>
</tbody>
</table>

"Unlawful" Civil Laws

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
<th>Maximum Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Federal</td>
<td>Unlawful direct disability discrimination</td>
<td>Section 5 of the Sex Discrimination Act 1984 (Cth)</td>
<td>For the purposes of this Act, a person (the discriminator) discriminates against another person (the aggrieved person) on the ground of a disability of the aggrieved person if, because of the disability, the discriminator treats, or proposes to treat, the aggrieved person less favourably than the discriminator would treat a person without the disability in circumstances that are not materially different.</td>
<td>Unlawful act. No criminal charge applicable.</td>
</tr>
<tr>
<td>NSW</td>
<td>Unlawful HIV/AIDS vilification</td>
<td>Section 49ZXB of the Anti-Discrimination Act 1977 (NSW)</td>
<td>It is unlawful for a person, by a public act, to incite hatred towards, serious contempt for, or severe ridicule of a person or group of persons on the ground that the person is, or members of the group are HIV/AIDS infected or thought to be HIV/AIDS infected (whether or not actually HIV/AIDS infected).</td>
<td>Unlawful act. No criminal charge applicable.</td>
</tr>
<tr>
<td>TAS</td>
<td>Unlawful inciting hatred</td>
<td>Section 19 of the Anti-Discrimination Act 1998 (Tas)</td>
<td>A person, by a public act, must not incite hatred towards, serious contempt for, or severe ridicule of, a person or a group of persons on the ground of any disability of the person or any member of the group.</td>
<td>Unlawful act. No criminal charge applicable.</td>
</tr>
</tbody>
</table>

No specific HIV/AIDS vilification laws in relation to online abuse in NT, QLD, SA, VIC and WA

HIV/AIDS vilification was considered illegal in the Australian Capital Territory only, with the maximum penalty consisting of a fine with no jail time. HIV/AIDS
vilification was considered an unlawful act to which no criminal charge is applicable in three jurisdictions. An additional five jurisdictions did not specify any related law. The data may indicate that Australian law considers HIV/AIDS vilification as a `less serious’ crime.

**Discussion**

Increasing media and scholarly attention has discussed the implications of sexual harassment online revealing women, along with the lesbian, gay, bisexual, transgender, gender diverse, intersex, queer, asexual and questioning community are the predominant targets (Powell & Henry, 2017). These claims were supported by interviews with ten Australian trolled journalists in chapter 4.1 (on p. 30). Although campaigns such as #MeToo have prioritised sexual harassment that occurs in person, when it occurs online, the applicability of existing legislation is altered, and this presents difficulties for law enforcement (Silver Sweeney, 2014; Citron, 2014). The emerging grey area has allowed abuse to flourish and its impacts on victims to be ignored (West, 2015; Williams; 2016). This study highlights the persistent and pervasive problem that inhibits online equality and suggests the scope of sexual harassment may be an indicator of a broader pattern of gender inequality and misogyny. As reflected by the feminist paradigm of this study, trolling behaviours are embedded in hegemonic masculine stereotypes and are a manifestation of gender power relations that serve to reinforce heterosexual and patriarchal ideologies.

An example arose in a 2004 case study when Sydney radio hosts John Laws and Steve Price were found to be in breach of homosexual vilification under the Anti-Discrimination Act 1997 (NSW). Listener Gary Burns had lodged a complaint after a discussion on radio 2UE about a gay couple on the television program The Block. This complaint process excludes the impact of offensive comments on members of the public who do not wish to pursue legal action at their own expense. This study suggests the laws should look at ways to punish vilification without the need for members of the public to personally initiate legal processes, a finding of this study also noted by Gelber and McNamara (2016). Price referred to the two men as “young poofs” and said that on the “uncut” version of The Block, “they could do
all sorts of grubby things at 11 o’clock at night.” Laws also referred to them as “a couple of young poofs” and said: “I don’t know what’s happened to Kerry [Packer]’s taste” (Burns v Radio 2UE Sydney Pty Ltd & Ors, 2004). The NSW Administrative Decisions Tribunal stated, “even if done in good faith and in the public interest [the comments were] not reasonable” and the pair were forced to issue an on-air apology, a written apology to be printed in *The Sydney Morning Herald* and pay Mr Burns legal costs. In addition, 2UE made a $10,000 donation to the HIV-AIDS charity the Bobby Goldsmith Foundation and promote the foundation with community service advertisements (AAP, 2008). However, payment of damages by way of compensation was not awarded to Mr Burns. The study highlights the onus on someone taking offense, rather the law working proactively, to curb such vilification, a significant legal issue for future research to extend on.

**Australia civil laws for personal injury claims against troll**

Personal injury damages include “death, pre-natal injury, impairment of a person's physical or mental condition and disease caused by the fault of another person” as defined in section 11 of the *Civil Liability Act 2002 (WA)*. Although no laws exist at federal level, Figures 55, 56, 57 and 58 describe Australian civil laws a journalist could implement to seek personal injury damages against a troll for online abuse.

**Figure 55**

*Australian civil laws for damaged for past economic loss that trolled journalists could use derived from chapter 4.7 data*

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>Damages for loss of earnings</td>
<td>Section 33 of the <em>Civil Law (Wrongs) Act 2002 (ACT)</em></td>
<td>In assessing damages for loss of earnings in relation to a claim, the court must disregard earnings above the limit of 3 times average weekly earnings a week. Loss of earnings means past economic loss because of loss of earnings or the deprivation or impairment of earning capacity.</td>
</tr>
<tr>
<td>NSW</td>
<td>Damages for past economic loss</td>
<td>Section 12 of the <em>Civil Liability Act 2002 (NSW)</em></td>
<td>This section applies to an award of damages for past economic loss due to loss of earnings or the deprivation or impairment of earning capacity, or In the case of any such award, the court is to disregard the amount (if any) by which the claimant’s gross weekly</td>
</tr>
</tbody>
</table>

*Chapter 4.7 Legal responses results and discussion* 337
<table>
<thead>
<tr>
<th>State</th>
<th>Damages for loss of earning capacity</th>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NT</td>
<td>Damages for loss of earning capacity</td>
<td>Section 20 of the <em>Personal Injuries (Liabilities and Damages) Act 2003 (NT)</em></td>
<td>In assessing damages to be awarded for: (a) past pecuniary loss due to loss of earnings or the deprivation or impairment of earning capacity; A court must disregard the amount (if any) by which the injured person’s gross weekly earnings would, but for the personal injury, have exceeded an amount that is 3 times average weekly earnings as published before 1 January preceding the date on which the assessment is made.</td>
</tr>
<tr>
<td>QLD</td>
<td>Damages for loss of earnings</td>
<td>Section 54 of the <em>Civil Liability Act 2003 (Qld)</em></td>
<td>In making an award of damages for loss of earnings, including in a dependency claim, the maximum award a court may make is for an amount equal to the limit. The limit is an amount equal to the present value of 3 times average weekly earnings per week for each week of the period of loss of earnings.</td>
</tr>
<tr>
<td>SA</td>
<td>Damages for loss of earning capacity</td>
<td>Section 54 of the <em>Civil Liability Act 1936 (SA)</em></td>
<td>If the injured person was incapacitated for work, damages for loss of earning capacity are not to be awarded in respect of the first week of the incapacity. Total damages for loss of earning capacity (excluding interest awarded on damages for any past loss) are not to exceed the prescribed maximum.</td>
</tr>
<tr>
<td>TAS</td>
<td>Damages for loss of earning capacity</td>
<td>Section 26 of the <em>Civil Liability Act 2002 (Tas)</em></td>
<td>Where a person is entitled to damages in respect of loss of earning capacity, a court must not award those damages on the basis the person was, or may have been capable of, earning income at greater than 3 times the adult average weekly earnings as last published by the Australian Bureau of Statistics before damages are awarded.</td>
</tr>
<tr>
<td>VIC</td>
<td>Damages for past economic loss</td>
<td>Section 28F of the <em>Wrongs Act 1958 (Vic)</em></td>
<td>This section applies to an award of damages for past economic loss due to loss of earnings or the deprivation or impairment of earning capacity. The maximum amount of damages that may be awarded for each week of the period of loss of earnings is an amount that is 3 times the amount of average weekly earnings at the date of the award.</td>
</tr>
<tr>
<td>WA</td>
<td>Damages for loss of earnings</td>
<td>Section 11 of the <em>Civil Liability Act 2002 (WA)</em></td>
<td>In assessing damages for loss of earnings, the court is to disregard earnings lost to the extent that they would have accrued at a rate of more than 3 times the average weekly earnings at the date of the award. Loss of earnings means past economic loss due to loss of earnings or the deprivation or impairment of earning capacity.</td>
</tr>
</tbody>
</table>

There are no specific federal laws for damages for past economic loss.
### Figure 56

*Australian civil laws for damaged for future economic loss that trolled journalists could use*

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>Damages for loss of earnings</td>
<td>Section 33 of the Civil Law (Wrongs) Act 2002 (ACT)</td>
<td>In assessing damages for loss of earnings in relation to a claim, the court must disregard earnings above the limit of 3 times average weekly earnings a week. Loss of earnings means future economic loss because of loss of earnings or the deprivation or impairment of earning capacity.</td>
</tr>
<tr>
<td>NSW</td>
<td>Damages for future economic loss</td>
<td>Section 12 of the Civil Liability Act 2002 (NSW)</td>
<td>This section applies to an award of damages: (a) for future economic loss due to the deprivation or impairment of earning capacity, or (b) for the loss of expectation of financial support. In the case of any such award, the court is to disregard the amount (if any) by which the claimant’s gross weekly earnings would (but for the injury or death) have exceeded an amount that is 3 times the amount of average weekly earnings at the date of the award.</td>
</tr>
<tr>
<td>NT</td>
<td>Damages for loss of earning capacity</td>
<td>Section 20 of the Personal Injuries (Liabilities and Damages) Act 2003 (NT)</td>
<td>In assessing damages to be awarded for: (a) future pecuniary loss due to the deprivation or impairment of earning capacity; or (b) the loss of expectation of financial support, A court must disregard the amount (if any) by which the injured person’s gross weekly earnings would, but for the personal injury, have exceeded an amount that is 3 times average weekly earnings as published before 1 January preceding the date on which the assessment is made.</td>
</tr>
<tr>
<td>NT</td>
<td>Future pecuniary loss</td>
<td>Section 21 of the Personal Injuries (Liabilities and Damages) Act 2003 (NT)</td>
<td>A court may award damages for future pecuniary loss only if satisfied by the claimant that the assumptions about the injured person’s future earning capacity, or the occurrence of other events on which the award is to be based, accord with the injured person’s most likely future circumstances had the personal injury not occurred. If a court is satisfied under subsection (1) about the claimant’s assumptions, it must adjust the amount of damages for future pecuniary loss (as assessed on those assumptions) by reference to the percentage possibility that the events might have occurred regardless of the personal injury. In awarding damages for future pecuniary loss, the court must state the assumptions on which the award is based and the relevant percentage by which damages have been adjusted.</td>
</tr>
<tr>
<td>QLD</td>
<td>Calculating present value of future loss</td>
<td>Section 57 of the Civil Liability Act 2003 (Qld)</td>
<td>When assessing an amount of damages as a lump sum for a future loss or gratuitous services, the amount must be the present value, calculated using the prescribed discount rate, of the future loss or gratuitous services.</td>
</tr>
<tr>
<td>SA</td>
<td>Lump sum compensation for future losses</td>
<td>Section 55 of the Civil Liability Act 1936 (SA)</td>
<td>(a) an injured person is to be compensated by way of lump sum for loss of future earnings or other future losses; and (b) an actuarial multiplier is used for the purpose of calculating the present value of the future losses, then, in determining the actuarial multiplier, a prescribed discount rate is to be applied.</td>
</tr>
</tbody>
</table>
TAS
Damages for loss of earning capacity
Section 28A of the Civil Liability Act 2002 (Tas)
If an award of damages is to include any component assessed as a lump sum for future loss, the present value of that future loss is to be qualified by adopting (a) a discount rate of 5 per cent; or (b) if another discount rate is prescribed, that other discount rate.

VIC
Damages for future economic loss
Section 28F of the Wrongs Act 1958 (Vic)
This section applies to an award of damages: (a) for future economic loss due to the deprivation or impairment of earning capacity; or (b) for the loss of expectation of financial support. The maximum amount of damages that may be awarded for each week of the period of loss of earnings is an amount that is 3 times the amount of average weekly earnings at the date of the award.

WA
Damages for loss of earnings
Section 11 of the Civil Liability Act 2002 (WA)
In assessing damages for loss of earnings, the court is to disregard earnings lost to the extent that they would have accrued at a rate of more than 3 times the average weekly earnings at the date of the award. Loss of earnings means future economic loss due to loss of prospective earnings or the deprivation or impairment of prospective earning capacity.

There are no specific federal laws for damages for future economic loss.

Figure 57

Australian civil laws for damaged for superannuation loss that trolled journalists could use derived from chapter 4.7 data

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSW</td>
<td>Damages for loss of superannuation entitlements</td>
<td>Section 15C of the Civil Liability Act 2002 (NSW)</td>
<td>The maximum amount of damages that may be awarded for economic loss due to the loss of employer superannuation contributions is the relevant percentage of damages payable (in accordance with this Part) for the deprivation or impairment of the earning capacity on which the entitlement to those contributions is based. The relevant percentage is the percentage of earnings that is the minimum percentage required by law to be paid as employer superannuation contributions.</td>
</tr>
<tr>
<td>QLD</td>
<td>Damages for loss of superannuation entitlements</td>
<td>Section 56 of the Civil Liability Act 2003 (Qld)</td>
<td>The maximum amount of damages that may be awarded to an employee for economic loss due to the loss of employer superannuation contributions is the relevant percentage of damages payable (in accordance with this part) for the deprivation or impairment of the earning capacity on which the entitlement to the contributions is based. The relevant percentage is the percentage of earnings that is the minimum percentage required by a written law to be paid on the employee’s behalf as employer superannuation contributions.</td>
</tr>
<tr>
<td>TAS</td>
<td>Damages for loss of superannuation entitlements</td>
<td>Section 25 of the Civil Liability Act 2002 (Tas)</td>
<td>The maximum amount of damages that may be awarded for economic loss due to the loss of employer superannuation contributions is the relevant percentage of damages payable for the deprivation or impairment of the earning capacity on which the entitlement to those contributions is based.</td>
</tr>
</tbody>
</table>
There are no specific federal laws or state laws in ACT, NT, SA, VIC, and WA for damages for superannuation loss.

Figure 58

*Australian civil laws for damaged for non-economic loss that trolled journalists could use derived from chapter 4.7 data*

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
<th>Cap on damages (as of August 2021)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>Damages for non-economic loss</td>
<td>Section 99 of the <em>Civil Law (Wrongs) Act 2002 (ACT)</em></td>
<td>In deciding damages for non-economic loss, a court may refer to earlier decisions of that or other courts for the purpose of establishing the appropriate award in the proceeding. For that purpose, the parties to the proceeding or their lawyers may bring the court’s attention to awards of damages for non-economic loss in those earlier decisions. This section does not change the rules for deciding other damages. In this section: non-economic loss includes the following: (a) pain and suffering (b) loss of amenities of life (c) loss of expectation of life</td>
<td>$350,000</td>
</tr>
<tr>
<td>NSW</td>
<td>Damages for non-economic loss</td>
<td>Section 16 of the <em>Civil Liability Act 2002 (NSW)</em></td>
<td>No damages may be awarded for non-economic loss unless the severity of the non-economic loss is at least 15% of a most extreme case. The maximum amount of damages may be awarded only in a most extreme case.</td>
<td>$687,000</td>
</tr>
<tr>
<td>NT</td>
<td>Damages for non-pecuniary loss</td>
<td>Section 27 of the <em>Personal Injuries (Liabilities and Damages) Act 2003 (NT)</em></td>
<td>A court must not award damages for non-pecuniary loss if the court determines the degree of permanent impairment to be less than 5% of the whole person. When awarding damages for non-pecuniary loss, a court must award the following amount: (a) if the court determines the degree of permanent impairment to be 85% or more of the whole person – the maximum amount: (b) if the court determines the degree of permanent impairment to be not less than 15% and not more than 84% of the whole person – the relevant percentage of the maximum amount on a sliding scale.</td>
<td>$432,500</td>
</tr>
<tr>
<td>QLD</td>
<td>Calculating general damages</td>
<td>Section 62 of the <em>Civil Liability Act 2003 (Qld)</em></td>
<td>For an injury arising after 1 December 2002, general damages must be calculated by reference to the general damages calculation provisions applying to the period within which the injury arose.</td>
<td>No maximum</td>
</tr>
</tbody>
</table>
### SA
| Damages for mental harm | Section 53 of the *Civil Liability Act 1936 (SA)* | Damages may only be awarded for mental harm if the injured person (a) was physically injured in the accident or was present at the scene of the accident when the accident occurred; or (b) is a parent, spouse, domestic partner or child of a person killed, injured or endangered in the accident. Damages may only be awarded for pure mental harm if the harm consists of a recognised psychiatric illness. Damages may only be awarded for economic loss resulting from consequential mental harm if the harm consists of a recognised psychiatric illness. | No maximum |

### TAS
| Damages for non-economic loss (general damages) | Section 28 of the *Civil Liability Act 2002 (Tas)* | In determining damages for non-economic loss, a court may refer to earlier decisions of that or other courts for the purpose of establishing the appropriate award in the proceedings. For that purpose, the parties to the proceedings or their counsel may bring the court’s attention to awards of damages for non-economic loss in those earlier decisions. | $350,000 |

### VIC
| Fixing damages for non-economic loss | Section 28G of the *Wrongs Act 1958 (Vic)* | The maximum amount of damages that may be awarded to a claimant for non-economic loss is $577,050. | $577,050 |

### WA
| Damages for non-pecuniary loss (general damages) | Section 9 of the *Civil Liability Act 2002 (WA)* | If the amount of non-pecuniary loss is assessed to be not more than Amount A for the year in which the amount is assessed, no damages are to be awarded for non-pecuniary loss. If the amount of non-pecuniary loss is assessed to be more than Amount A but not more than Amount C for the year in which the amount is assessed, damages for non-pecuniary loss are not to be awarded in an amount that is more than the excess of the amount assessed over Amount A. Non-pecuniary loss means: (a) pain and suffering; and (b) loss of amenities of life; and (c) loss of enjoyment of life; and (d) curtailment of expectation of life; and (e) bodily or mental harm. | $407,500 |

There are no specific federal laws for damages for non-economic loss.

The study identified the cap for damages for non-economic loss in six jurisdictions. New South Wales had the highest cap of $687,000 and the Australian Capital Territory and Tasmania had the lowest, set at $350,000. The caps increase over time and the figure is set by the Attorney General of each state or territory; however, there are some provisions for aggravated damages, which can be higher.
Queensland and South Australia do not have maximum damages and provisions are calculated applying to the period within which the injury arose. The universal adoption of this maximum amount may suggest a national consensus about the attributed severity of personal injury claims.

Legal debate about personal injury claims for online abuse continues to play out in court. In a 2020 defamation case, Kellie Smith and Stacey Isaac were ordered to pay a combined total of $150,000 for ten social media posts they made about a wedding planner Tristan Moy. The courts awarded damages for three purposes: personal distress, injury to personal and business reputation, and vindication of reputation (Moy v Smith & Isaac, 2020). Media coverage of successful cases has begun to influence public understanding of the laws. However, the continued presence of trolling online suggests limited impact.

**Australian civil laws for duty of care against employer**

Under civil laws for personal injury, journalists can seek damages against their employer under the grounds the media organisation failed in their duty of care to intervene and protect the victim. These laws are described in Figure 59 below; however, no laws are specified at federal level.

**Figure 59**

*Australian civil laws for duty of care that trolled journalists could use derived from chapter 4.7 data*

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>Mental harm duty of care</td>
<td>Section 34 of the <em>Civil Law (Wrongs) Act 2002 (ACT)</em></td>
<td>A person (the defendant) does not owe a duty to another person (the plaintiff) to take care not to cause the plaintiff mental harm unless a reasonable person in the defendant’s position would have foreseen that a person of normal fortitude in the plaintiff’s position might, in the circumstances of the case, suffer a recognised psychiatric illness if reasonable care were not taken. For the application of this section in relation to pure mental harm to a person, the circumstances of the case to which the court must have regard include: (a) whether or not the mental harm was suffered as the result of a sudden shock; and (b) whether the plaintiff witnessed, at the scene, a person being killed, injured, or put in danger; and (c) the nature of the relationship between the plaintiff and anyone killed, injured, or put in danger; and (d) whether or not...</td>
</tr>
</tbody>
</table>
there was a pre-existing relationship between the plaintiff and the defendant.
For the application of this section in relation to consequential mental harm to a person, the circumstances of the case to which the court must have regard include the nature of the bodily injury out of which the mental harm arose.
This section does not affect the duty of care a person (the defendant) has to another person (the plaintiff) if the defendant knows, or ought reasonably to know, that the plaintiff is a person of less than normal fortitude.

| NSW | Duty of care | Section 5B of the *Civil Liability Act 2002 (NSW)* | A person is not negligent in failing to take precautions against a risk of harm unless:
(a) the risk was foreseeable (that is, it is a risk of which the person knew or ought to have known), and
(b) the risk was not insignificant, and
(c) in the circumstances, a reasonable person in the person's position would have taken those precautions.
In determining whether a reasonable person would have taken precautions against a risk of harm, the court is to consider the following (amongst other relevant things):
(a) the probability that the harm would occur if care were not taken,
(b) the likely seriousness of the harm,
(c) the burden of taking precautions to avoid the risk of harm,
(d) the social utility of the activity that creates the risk of harm. |
| QLD | Breach of duty | Section 9 of the *Civil Liability Act 2003 (Qld)* | A person does not breach a duty to take precautions against a risk of harm unless:
(a) the risk was foreseeable (that is, it is a risk of which the person knew or ought reasonably to have known); and
(b) the risk was not insignificant; and
(c) in the circumstances, a reasonable person in the position of the person would have taken the precautions.
In deciding whether a reasonable person would have taken precautions against a risk of harm, the court is to consider the following (among other relevant things):
(a) the probability that the harm would occur if care were not taken;
(b) the likely seriousness of the harm;
(c) the burden of taking precautions to avoid the risk of harm;
(d) the social utility of the activity that creates the risk of harm. |
| SA | Mental harm duty of care | Section 33 of the *Civil Liability Act 1936 (SA)* | A person (the defendant) does not owe a duty to another person (the plaintiff) to take care not to cause the plaintiff mental harm unless a reasonable person in the defendant's position would have foreseen that a person of normal fortitude in the plaintiff's position might, in the circumstances of the case, suffer a psychiatric illness.
For the purposes of this section:
(a) in a case of pure mental harm, the circumstances of the case to which the court is to have regard include the following:
(i) whether or not the mental harm was suffered as the result of a sudden shock;
(ii) whether the plaintiff witnessed, at the scene, a person being killed, injured or put in peril;
(iii) the nature of the relationship between the plaintiff and any person killed, injured or put in peril;
(iv) whether or not there was a pre-existing relationship between the plaintiff and the defendant; |
(b) in a case of consequential mental harm, the circumstances of the case include the nature of the bodily injury out of which the mental harm arose.
This section does not affect the duty of care of a person (the defendant) to another (the plaintiff) if the defendant knows, or ought reasonably to know, that the plaintiff is a person of less than normal fortitude.

<table>
<thead>
<tr>
<th>State</th>
<th>Breach of Duty</th>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAS</td>
<td>Breach of Duty</td>
<td>Section 11 of the Civil Liability Act 2002 (Tas)</td>
<td>A person does not breach a duty to take reasonable care unless: (a) there was a foreseeable risk of harm (that is, a risk of harm of which the person knew or ought reasonably to have known); and (b) the risk was not insignificant; and (c) in the circumstances, a reasonable person in the position of the person would have taken precautions to avoid the risk. In deciding whether a reasonable person would have taken precautions against a risk of harm, the court is to consider the following (among other relevant things): (a) the probability that the harm would occur if care were not taken; (b) the likely seriousness of the harm; (c) the burden of taking precautions to avoid the risk of harm; and (d) the potential net benefit of the activity that exposes others to the risk of harm.</td>
</tr>
<tr>
<td>TAS</td>
<td>Mental harm duty of care</td>
<td>34 of the Civil Liability Act 2002 (Tas)</td>
<td>A person (&quot;the defendant&quot;) does not owe a duty to another person (&quot;the plaintiff&quot;) to take care not to cause the plaintiff mental harm unless a reasonable person in the position of the defendant ought to have foreseen that a person of normal fortitude might, in the circumstances of the case, suffer a recognised psychiatric illness if reasonable care were not taken. For the purpose of the application of this section in respect of pure mental harm, the circumstances of the case include the following: (a) whether or not the mental harm was suffered as the result of a sudden shock; (b) whether or not there was a pre-existing relationship between the plaintiff and the defendant. For the purpose of the application of this section in respect of consequential mental harm, the circumstances of the case include the nature and extent of personal injury suffered by the plaintiff. This section does not require the court to disregard what the defendant knew or ought to have known about the fortitude of the plaintiff.</td>
</tr>
<tr>
<td>VIC</td>
<td>Duty of care</td>
<td>Section 28F of the Wrongs Act 1958 (Vic)</td>
<td>A person is not negligent in failing to take precautions against a risk of harm unless: (a) the risk was foreseeable (that is, it is a risk of which the person knew or ought to have known); and (b) the risk was not insignificant; and (c) in the circumstances, a reasonable person in the person's position would have taken those precautions. In determining whether a reasonable person would have taken precautions against a risk of harm, the court is to consider the following (amongst other relevant things): (a) the probability that the harm would occur if care were not taken; (b) the likely seriousness of the harm; (c) the burden of taking precautions to avoid the risk of harm; and (d) the social utility of the activity that creates the risk of harm.</td>
</tr>
</tbody>
</table>
WA Mental harm duty of care Section 55 of the Civil Liability Act 2002 (WA) A person (the defendant) does not owe a duty of care to another person (the plaintiff) to take care not to cause the plaintiff mental harm unless the defendant ought to have foreseen that a person of normal fortitude might, in the circumstances of the case, suffer a recognised psychiatric illness if reasonable care were not taken.
For the purpose of the application of this section in respect of pure mental harm, the circumstances of the case include the following:
(a) whether or not the mental harm was suffered as the result of a sudden shock;
(b) whether the plaintiff witnessed, at the scene, a person being killed, injured or put in peril;
(c) the nature of the relationship between the plaintiff and any person killed, injured or put in peril;
(d) whether or not there was a pre-existing relationship between the plaintiff and the defendant.
For the purpose of the application of this section in respect of consequential mental harm, the circumstances of the case include the personal injury suffered by the plaintiff.
This section does not require the court to disregard what the defendant knew or ought to have known about the fortitude of the plaintiff.

There are no specific federal laws or NT laws for damages for duty of care.

In addition to chapter 4.2's findings (on p. 54), academic research (Citron, 2014; Jane, 2015; Phillips, 2015) and mainstream media (Bartlett, 2013; Greig, 2016; Williams, 2016; Tuohy, 2019; Molan, 2020; Hyland, 2021) have now documented the emotional and psychological distress that online abuse has on all genders, presenting a serious workplace issue for Australian newsrooms. In-depth analysis of specific responses to trolling by The Sydney Morning Herald, News.com and ABC were discussed in chapter 4.6 (on p. 202); however, the slow implementation of strategies, and the absence of universal approach, suggests some Australian newsrooms could be exposed to potential duty of care claims.

Landmark case
One such example occurred on February 22, 2019, in the landmark case YZ (a pseudonym) v The Age Company Limited. The County Court of Victoria awarded $180,000 in damages to a former (unnamed) journalist at The Age. In a world first, the judge ruled the newsroom was responsible for the journalist's PTSD because it failed to provide a safe workplace (Edraki & Carrick, 2019). In damning witness evidence, the editorial training manager of The Age voiced frustration at being unable to persuade management to implement a suitable training and support program in comparison to the ABC, which had had a peer-support program in
place for a decade (Ricketson & Wake, 2019). Recommendations for media outlets to seriously consider risk management programs that include training programs, peer support and self-recognition symptoms are no longer novel; however, prioritising their value and implementing these strategies within Australian newsrooms is varied and discussed in chapter 4.6 (on p. 202).

**Post-Traumatic Stress**

Historically, the idea of journalists suing their employers for occupational post-traumatic stress was unheard of and newsroom culture dictated that journalists performed any task required of them without question (Ricketson & Wake, 2019). This newsroom culture developed as a key contributor to journalists feeling unable to express concerns about workplace trauma for fear of appearing unsuitable to the profession (Oakham, 2004; Ricketson & Wake, 2019). However, the seminal case emerged as a catalyst for widespread debate criticising the lack of accountability altered by media organisations for similar cases of psychiatric injury caused by online abuse endured in the workplace. The landmark ruling against *The Age* could have far-reaching implications for newsrooms and force them to reconsider their duty of care when it comes to trauma and mental health of employees (Edraki & Carrick, 2019). In another milestone case on October 23, 2019, Seven Network was forced to pay *House Rules* reality TV contestant Nicole Prince damages for psychological injury caused by the online abuse she received after appearing on the show. The constant barrage including threats of physical assault made her fearful for her life, compromised her personal safety and obstructed her ability to obtain work (Chung, 2019). Although the precedent has been set for journalists to sue employers for trolling-related post-traumatic stress, there is no evidence of widespread use of this pathway in the current media landscape (as at June 2021), but this does not rule out the existence of such cases. Executive director of the Dart Centre for Journalism and Trauma Bruce Sharpio suggests they “usually get settled out of court” (Sharpio, as cited in Edraki & Carrick, para. 18) and are therefore not accessible in retrievable archival evidence.

**Systems approach is required**

While giving evidence at the 2018 hearing of the Australian Senate enquiring into
the adequacy of existing offences in the Commonwealth Criminal Code and of state and territory criminal laws to capture cyberbullying, law firm Maurice Blackburn argued in order to protect journalists who are required to communicate online as part of their work, there should be: “enforceable sanctions against employers who fail in the duty of care to provide a safe workplace for their employees” (as cited in Gorman, 2019, p. 81). However, employment lawyer, Josh Bornstein proposed the responsibility for maintaining a safe social media space “should probably not fall solely on employers” (Bornstein, as cited in Gorman, 2019, p. 71), which resonates with a multilayered approach to the workplace health and safety issue of cyberhate in support of Ackoff’s (1974) Systems Theory. Although research by Posetti in 2018 had established training modules for larger news organisation such as Nine (formerly Fairfax Media), Nine refused to provide a copy of for analysis in this study. As Posetti was unavailable for an interview during the duration of this study, a booklet funded by United Nations Educations Scientific and Cultural Organisation’s (UNESCO) and made by Ireton and Posetti (2018) called, Journalism, ‘fake news’ and disinformation: A handbook for journalism education and training, was obtained for analysis. However, as previously discussed (on p. 238), the information provided was vague, did not provide relevant workplace training and was of minimal use for employers and working journalists in Australia. Specific guidelines determining best-practice remain a widely debated and not yet adequately resolved issue of the Australian contemporary media landscape. Duty of care legislation and workers’ compensation are interwoven; however, duty of care claims must prove the employer is at fault.

**Australian civil laws for workers compensation**

Workers’ compensation is legislation that provides financial compensation for employees injured at work (Oxford English Dictionary, 2021) to cover employee wages while they are not fit for work, in addition to medical expenses and rehabilitation (Fair Work Australia, 2021). Australian workers’ compensation and injury management schemes are based on a ‘no-fault’ principle, which means trolled journalists who suffered a psychological harm do not have to establish that their employer was at fault or negligent to make a claim (Work Cover WA, 2021).
Under civil laws for personal injury, journalists can also seek damages from employees and insurance companies under workers compensation. Due to the length of these documents, individual descriptions were excluded from the data set. The acts for each jurisdiction are listed in the following table, however, there are no acts specified at federal level or in the Northern Territory.

**Figure 60**

*Australian civil laws worker’s compensation that trolled journalists could use*

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>Workers’ Compensation</td>
<td>Section 31 of the <em>Workers Compensation Act 1951 (ACT)</em></td>
</tr>
<tr>
<td>NSW</td>
<td>Workers’ Compensation</td>
<td>Section 149 of the <em>Workers Compensation Act 1987 (NSW)</em></td>
</tr>
<tr>
<td>QLD</td>
<td>Workers’ Compensation</td>
<td>Section 33 of the <em>Workers’ Compensation and Rehabilitation Act 2003 (Qld)</em></td>
</tr>
<tr>
<td>SA</td>
<td>Workers’ Compensation</td>
<td>Section 30 of the <em>Return to Work Act 2014 (SA)</em></td>
</tr>
<tr>
<td>TAS</td>
<td>Workers’ Compensation</td>
<td>Section 32 of the <em>Workers Rehabilitation and Compensation Act 1988 (Tas)</em></td>
</tr>
<tr>
<td>VIC</td>
<td>Workers’ Compensation</td>
<td>Section 8 of the <em>Workers Compensation Act 1958 (Vic)</em></td>
</tr>
<tr>
<td>WA</td>
<td>Workers’ Compensation</td>
<td>Section 18 of the <em>Workers’ Compensation and Injury Management Act 1981(WA)</em></td>
</tr>
</tbody>
</table>

There are no specific federal or NT laws for workers compensation.

There are no studies that confirm the most effective strategies for managing online harassment (Slaughter & Newman, 2020) and proposed anti-trolling laws in 2022 are debated and further criticised to potentially “worsen online abuse” (Karp, 2022, para. 1). However, failure to protect journalists from abuse resulting in personal injury provides the grounds for workers’ compensation claims. As outlined by the Dart Centre for Journalism and Trauma, several claims have been successfully prosecuted on the basis, “the worker failed to establish that the risk of suffering a recognisable psychiatric was reasonably foreseeable” by the employer (Knoester, 2014, para. 14). This causative link was first brought to light in the seminal study by psychologist Cait McMahon (1993) who found not only do journalists report quite serious trauma at the time of covering stories of a critical nature, but they also continued to experience residual effects afterwards, with some of the symptoms in the realm of post-traumatic stress and acute stress disorders. McMahon went on to establish the Australasian branch of the Dart...
Centre for Journalism and Trauma six years later in 1999, compiling academic research to cultivate a more universal acceptance about the potential impacts of post-traumatic stress within the contemporary journalism community. Growing documented evidence soon warranted consideration in Australian courtrooms, and while the courts commonly deal with such claims, it was not until February 22, 2019, in the previously discussed case of YZ (a pseudonym) v The Age Company Limited, that journalists were considered to fall within this scope. Although the case was not about trolling and was not the first time a journalist had sued over occupational post-traumatic stress, it was the first time one had succeeded and set the precedent for future verdicts (Edraki & Carrick, 2019). Common-law claims for psychiatric injury suffered in the workplace are often the most complex and difficult matters for personal injury practitioners (Knoester, 2014).

**Civil laws for defamation**

Prior to the internet, the impact of defamation law was largely limited to national borders; therefore, conflicts that arose as a result of different cultural interpretations of rights and responsibilities in relation to freedom of expression were uncommon (Dare, 2005). Cross-cultural disputes of this nature were most often articulated through defamation suits between traditional publishers, with the plaintiffs most often being wealthy individuals or corporations (Dare, 2005). However, the potential for cross-cultural disputes has increased dramatically with the rapid expansion of the internet. Part of growing up is learning to take responsibility for comments that are hurtful: defamation law reflects that lesson (Worrall, 2020). Discussions about how to best define and implement defamation laws in a journalism workplace, interwoven with active online dialogue, continues in mainstream media and extends into current academic debates in 2021. Each Australian state or territory has their own respective defamation act that depicts what constitutes defamation and although there is no specific federal law, in 2005, the unification of state laws synchronised responses from all jurisdictions. Due to the length of the currently used 2005 defamation legislation, individual descriptions were excluded from the data set in Figure 61 below, but in summary govern “the wrong of injuring another’s reputation without good reason or justification” (p. 219) as defined by Pearson and Polden (2019). As an example,
the case study *Cullen v White 2005* is further explored in the discussion that follows. The Media Entertainment and Arts Alliance voiced grave concerns about the mounting necessity for defamation law reform in 2018 and, given the vast changes in technology in the 16 years since 2005, continued pressure from multiple stakeholders saw reform as long overdue (Todd et al, 2020), and in 2020 commenced amendments to the model defamation provisions. The defamation reformation ensured “Australia’s laws received a makeover for the digital age” (Shaw, 2021). The study therefore notes the following 2005 laws will be updated in the near future and are currently in their second stage of review as of August 2021.

**Figure 61**

*Australian civil laws for defamation that trolled journalists could use*

<table>
<thead>
<tr>
<th>Jurisdiction</th>
<th>Law</th>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT</td>
<td>Defamation</td>
<td>Defamation Act 2001 (ACT)</td>
</tr>
<tr>
<td>NSW</td>
<td>Defamation</td>
<td>Defamation Act 2005 (NSW)</td>
</tr>
<tr>
<td>NT</td>
<td>Defamation</td>
<td>Defamation Act of 2006 (NT)</td>
</tr>
<tr>
<td>QLD</td>
<td>Defamation</td>
<td>Defamation Act 2005 (Qld)</td>
</tr>
<tr>
<td>SA</td>
<td>Defamation</td>
<td>Defamation Act 2005 (SA)</td>
</tr>
<tr>
<td>TAS</td>
<td>Defamation</td>
<td>Defamation Act 2005 (Tas)</td>
</tr>
<tr>
<td>VIC</td>
<td>Defamation</td>
<td>Defamation Act 2005 (Vic)</td>
</tr>
<tr>
<td>WA</td>
<td>Defamation</td>
<td>Defamation Act 2005 (WA)</td>
</tr>
</tbody>
</table>

There are no specific federal laws for defamation.

Defamation in this context is the legal term for reputational damage that can result from trolling attacks about the character and activities of a journalist. However, it is more commonly committed by journalists. Abusive commentary that emerged on comment streams below news articles and on social media platforms such as Facebook and Twitter can easily be accessed by the public and has the potential to tarnish a journalist’s reputation, impacting both professional and social realms. The instantaneous nature of the internet does not require truth to be established before publishing, facilitating the swift spread of deceptive comments that can
inflict damage. Although specific remarks may be removed, the accrued harm may linger and impede a journalist’s employment potential. Examples had become so widespread the term “Twibel” was coined by blending Twitter and libel in 2014 (Placid, Wynekoop & Feicht, 2016).

**Defamation case study – Professor Trevor Cullen**

A notable example is that of Professor Trevor Cullen, a supervisor of this study, whose case sparked interest in trolling as an investigative topic. Cullen became the target of a sustained campaign of online harassment and defamation by United States resident Bill White, with a range of fabricated accusations of paedophilia and academic fraud. On September 3rd, 2003, the Supreme Court ruled the conduct of the defendant could only be attributed to a conscious desire to cause the maximum amount of damage, hurt and embarrassment to the plaintiff, by what amounted to a campaign of deliberately offensive vilification (Dare, 2005). Cullen was awarded $95,000 in damages (*Cullen v White* 2003); however, difficulties in registering the judgement in the United States meant the judgement could not be satisfied, with Cullen quoted US$350,000 with no assurance of success to pursue the damages in the United States. Despite the verdict, the residual emotional impact remained. As highlighted on the criminal law flowchart, problems arise with the legal process when the abuser does not live in the same country as the victim. This is common to incidents that emerge in online spaces; however, only sporadic cases have been successful including *Dow Jones & Company Inc v Gutnick* on December 10, 2002. Reactions from media and technology groups after the ruling were particularly critical, suggesting the decision represented a threat to freedom of expression on the Internet and a deterrent to online publication (Garnett, 2003). However, the landmark case represented the first major opportunity for an Australian court to examine the application of the rules of private international law to Internet conduct.

**Chilling effect of defamation laws**

Although not directly related to its scope of trolled journalists, the study acknowledged an emerging problem between defamation law and news reporting that was unpacked for a brief discussion. Labelled as a war on journalism (Media
Entertainment Arts Alliance, 2018), sentiments suggested inadequate legislation has suppressed press freedom in Australia. However, journalists who resort to this law to prosecute trolls could spark an interesting debate on their relationship to audience critique. Although the laws are under review in 2021, current defamation action against media organisations often requires substantial legal expenses to protect against potential damages. These costs further strain media organisations already burdened with profitability struggles in the wake of digital disruptions on traditional revenue models. Along with this economic impact, defamation laws had a chilling effect on reporting in Australia (Taylor & Karp, 2019), and can intimidate journalists from covering legitimate news stories in the public interest and applying scrutiny to the rich and powerful because of fears their work may result in costly and lengthy litigation. When the risk of defamation law can be leveraged to muzzle the media in such a way, both democracy and press freedom are jeopardised. This would make it harder to pursue defamation.

Defamation against media organisations for trolling

Debate arose within mainstream media and academia about whether websites and online platforms should be considered publishers, and therefore be held liable for the abuse they host. A landmark case identified by the study is that of Voller v Fairfax Media, Nationwide News and Sky News (2019) NSWSC 766, which demonstrated publishers could be held liable for abusive comments made by a third party. On June 24, 2019, former Northern Territory youth detainee Dylan Voller won a defamation case against three of Australia’s largest media organisations Fairfax Media, Nationwide News and Sky News after abusive comments were made on the Facebook pages of The Sydney Morning Herald, The Australian, Sky News, The Bolt Report and The Centralian Advocate. The court upheld claims that the newsrooms should have recognised the significant risk of defamatory observations after publishing the article and failed to effectively monitor or hide malicious commentary (McKinnell, 2019). Occasional comment moderation was not considered sufficient, particularly for media organisations of significant size with extensive resources, and strategies should have been implemented by the newsrooms to ensure a more robust and deliberate attempt at monitoring discourse (Taylor, 2020). With a final dismissal by the New South
Wales Court of Appeal on June 1, 2020, the decision set a new precedent that had wide-ranging ramifications for Australian media companies, which could now be held responsible for the comments of users who engage with the content they produce (Worrall, 2020). The Court of Appeal’s majority ruling explained, “defamation is an actionable wrong that lies in the publication to a reader, listener or observer of matter that injures another person’s reputation,” and does not require intention. Critics suggested the verdict “stretched common-sense intuition” (Worrall, 2020, para. 3) and would require newsrooms to heavily invest in more comment moderation. Left exposed under the new ruling, the verdict opened the avenue for trolled journalists to sue media organisations that encouraged user engagement resulting in defamatory trolling content. Media companies were not just responsible for the content written by their journalists but were now considered the ‘publishers’ of comments made by readers on their social media accounts and could be sued for such comments.

**Defamation against online platforms for trolling**

Debate arose in mainstream media (Kiss & Arthur, 2013; Bauckhage, 2014) and academia (Laurenson, 2015) about whether online platforms should be considered publishers and therefore be held liable for the abuse they host. Although Google voiced concerns and argued search engines should not be required to be gatekeepers of defamatory content (Taylor & Karp, 2019), the following cases illustrate they can be held liable under Australian law. A key case setting precedence for corporations to be held liable for defamation claims is the 2015 case of *Duffy v Google Inc*. In the landmark trial, South Australian Dr Janice Duffy was awarded $100,000 plus interest in general damages, a decision upheld after appeal (*Google Inc v Duffy*, 2017). This case marked a new legal era, one that held publishers accountable for defamatory material made by others on their platform. Although a subordinate publisher, Google was considered liable, a verdict that opened pathways for defamation cases against social media platforms, moving discussions of liability back into the media spotlight. Facilitating the publication of defamatory commentary may now constitute as participation providing the grounds for prosecution under Australian civil law. More recently in the case of *Defteros v Google 2020*, Google was ordered to pay...
$40,000 in damages to a Melbourne lawyer after a Supreme Court of Victoria ruling found the internet giant was a ‘publisher’ and had defamed the man.

**Defamation against social media companies for trolling**

Social media platforms were not considerable factors when the eight state and territories unified their defamation legislation in 2005 as Twitter did not exist at the time, and Facebook was in its early stages and not yet popular in Australia in its first two years of operation. Although issues of cyberbullying and workplace harassment in Australia were high in the public consciousness in 2021 (Diaz, 2021), there is no universal agreement about the accountability for social media companies that host such abuse and debates about which stakeholders should be held accountable for trolling abuse continues to develop in court. The study identified two key Australian lawsuits that set the precedent for social media companies to be held liable for abusive content on their platforms. The Meggit v Hardy 2012 case demonstrated that an international social media corporation could be prosecuted under Australian legislation (Black, 2014). This landmark case was the first under Australian law to see a social media company sued for defamation. In February 2012, Joshua Meggit commenced a lawsuit against Twitter after being wrongly identified as the author of a hate blog directed at writer and TV personality Marieke Hardy. Although Meggitt and Hardy reached a confidential legal settlement (Black, 2014), the broader issue of Twitter’s moral culpability and legal responsibility for defamatory tweets began to emerge in public discourse. Precedents set by the Gutnick v Dow Jones 2002 case held that a cause of action for defamation arises in all places where the story can be read, permitting international corporations to be prosecuted in Australian courts. Therefore, Twitter could be held liable in Australia for publication of tweets on its platform, although the company is based in the United States. The case highlighted that although disclaimers in the terms and conditions offer some legal protection, they do not provide absolute immunity. This case marked one of the first of its kind to hold a platform liable for abusive commentary, rather than an individual perpetrator. The study notes international laws vary, and particularly differ in the United States. Section 230 of their Communications Decency Act provides immunity from liability for providers and users of an "interactive computer
service” who publish information provided by others. However, in Australia, there is no such equivalent exemption, and social media platforms must rely on “innocent dissemination” as a defence or clause 91 of Schedule 5 of the Broadcasting Services Act 1992, which provides a general limitation of liability for internet service providers and content hosts from Australian State and Territory laws for material hosted on behalf of another. While these provisions clearly apply to internet service providers, their extension to provide immunity for social media platforms was considered unlikely (Black, 2014). Academic and lawyer, Simon Black, was a vocal critic in 2014 and suggested that online intermediaries should not be held liable for the action of their users, and further legal reform was required to protect social media hosts. In further criticism of the practicality of such a responsibility, Black (2014) argued platforms would either be forced to pass on the considerable cost of monitoring to the public or withdraw from Australia altogether and warned the potential punishment of social media companies for abusive content may act as a powerful disincentive to innovative new services being developed and based in Australia. More balanced suggestions propose legislature should only impose additional levels of liability upon online intermediaries after they are advised of defamatory comments and fail to respond (Black, 2014). Although Australian law had detailed provisions for offers to make amends for defamation, legislation struggled to keep up with rapid technological advances over the past few decades; however, the Gutnick v Dow Jones 2002 case exposed, and the Meggit v Hardy 2012 case solidified, the potential strength of Australia’s existing defamation law with respect to online intermediaries and platforms.

**Calls for Australian law reform in response to trolling**

As most Australian legislation was written before the advent of the internet, criticisms about of its effectiveness in contemporary journalism practice often make comparisons to other countries. While Australian laws predominantly focus on physical violence, German laws frame hate speech and online defamation as serious crimes that require hosts to remove content within 24 hours of notification or face fines of 50 million euro (Lomas, 2020). Although the study notes this is not a perfect resolution, the strategy emerged from more nuanced
conversations and coherent policy positions than the current Australian legislation. Consistent external pressures from campaigning and widespread media coverage compelled Australian Prime Minister Scott Morrison to promise an increased maximum penalty for trolling in 2019 if re-elected (Murphy, 2019). This defamation law bill supported actions by the like of Australian Minister for Defence Peter Dutton against refugee activist Shane Bazzi over tweet that labelled him a ‘rape apologist’ (Karp, 2021, November). Inclusion as a central electoral strategy illustrated the understanding and importance of anti-trolling policies within public opinion at the time. Commenting on the announcement, interviewee for this study Dr Jennifer Beckett (2019) argued there seemed to be “little understanding” within the Coalition and across the political spectrum “about the impact of online abuse in the real world and how to effectively regulate social media platforms”. Although Facebook CEO Mark Zuckerberg himself called for regulation, Beckett (2019) stressed the need for government vigilance to not fall into the trap of regulating social media providers in ways they want to be regulated. Journalists require good policy from politicians from which to create informed and constructive legislation. As part of a systems approach (Ackoff, 1974), a commitment to international cooperation on global policy is a critical requirement of future law reform, in addition to appropriate resourcing and training for those who deal with the consequences of platform misgovernance including police, social workers and educators. Although legislative promises provided some comfort for trolling victims, the mooted policies are still being discussed as of August 2021. After years of calls from activists, legislative reform has begun to shift the focus from coping mechanisms for victims to punishments for perpetrators.

**Proposed new Australian laws in response to trolling**

The heightened tenor, tone and velocity of trolling pressured the government to respond after the number of reports submitted to the eSafety Commission surpassed youth cyber-bullying reports over the Covid-19 period (Dagg, 2021). In a historical legislative and social milestone, the Australian Government committed $39.4 million to a cyber abuse takedown scheme on October 12, 2020 (Dudley-Nicholson, 2020) and announced it would introduce tough new laws in response
to trolling on December 23, 2020 (Davey, 2020). The landmark *Online Safety Bill 2021* was presented to parliament in February 2021 and was heralded as the world's first cyber abuse takedown laws (Visentin, 2020). Although the laws have not yet passed in August 2021, cross-party support suggest they are likely to be passed by the conclusion of 2021 (Wilson, 2021). Individuals who abuse, threaten, intimidate, or vilify others online would now face fines of up to $111,000 (Molan, 2020) and could "lose their home" (Tamworth, 2021, para. 11) if they are unable to pay. If the bill passes corporations such as Facebook and Twitter who refuse to take down offensive material or reveal the identities of perpetrators face fines of $555,000 (Davey, 2020). Social networks will have 24 hours to remove vile, menacing messages reported by the Australian eSafety Commissioner, a position held in 2021 by Julie Inman Grant (Dudley-Nicholson, 2020). The Bill also provides more power to unmask internet trolls, with Grant (2020) commenting, "[as] we know that anonymity is often a facet of trolling [these laws] will allow us to compel the social media sites to give us device information and other information so we can track down the identity of the troll" (Grant, as cited in Macmillan, 2020, para. 10). It was vital that Australians now had an advocate with powers to force social networks to remove online abuse, as the tech firms were often slow to address harmful online content. Supporters suggest that the online disinhibition effect will be hindered once the assurance of anonymity is removed, and the potential of real-life consequences are more widely understood. As suggested in *Panic* by Australian journalist and author David Marr (2013), a fear of harsh penalties can act as an effective deterrent for negative behaviours. Although legislation is clearer about the illegality of abusive conduct in the real world, public discourse suggests that such behaviours online should likewise be criminal (Molan, 2020). Australian journalist and advocate for the new laws Erin Molan (2020) highlighted society now moves seamlessly between both worlds suggesting “what we deem acceptable behaviour in the real world needs to be reflected in the online space” (para. 13). Prior coping strategies suggested victims abstain from interacting online; however, the new laws express reviewed sentiments that demand the perpetrators be the ones removed from the platforms. Critical debate proposed the need for balanced legislation that considered concerns about potential restrictions on freedom of speech (Tamworth, 2021). However, advocates
strongly opposed suggestions that the proposed laws would stifle free expression (Molan, 2020), and stressed the laws would not prevent users from disagreeing, expressing opinions, being highly critical of others or just being mean; more accurately, the laws would punish serious offences. Unbridled free speech, including hate speech, threats, racism and misogyny, can discourage participation and lead to suppression of speech as the voices of victims can be left unprotected. This study recognises the inherent difficulty in striking a balance between enabling freedom of expression and protecting people from the harms associated with online trolling. The dilemma is in determining this optimum level of regulation without undermining the principles of free speech. Social media often brings to the surface the reality of the human condition, and this ugliness expressed online must be addressed. To tackle such a social issue, the law requires higher standards for digital civility. Abuse can have devastating impacts on victims, leading to lingering emotional and mental distress that the new proposed laws begin to address more adequately.

**Summary**

New legislation may be part of the solution, but it is not an effective standalone solution. Although the current harassment law of “using a carriage service to menace or offend” has been around for 26 years, both the vast number of victims and the sustained and endemic presence of abusive commentary online suggests the current legislation is not an effective deterrence strategy. However, layering additional legislation on top of existing laws may not be the sole solution, and the problem needs to be addressed through a holistic lens. This chapter aimed to provide a clear pathway for journalists who wish to commence a legal process against a troll. While victims can report incidences, criminal legislation is only as effective as the policing of the existing laws. A critical part of the problem is the knowledge of the application of existing laws, and responsiveness of law enforcement officers. The study suggests further resources may need to be allocated to police for further education and training about how to investigate and prosecute online offenders. The study also recommends that specified legislative pathways for trolling be provided to Australian journalists by newsrooms and educational institutions, such as those outlined in this chapter, to aid navigation.
of the difficult legal process.

This chapter analysed relevant Australian laws that a trolled journalist could use to prosecute a troll. Legislation and legal action sit at the governance level of the systems model, but in order to be effective, laws need to be understood, enforced, and used by all stakeholders in the trolling paradigm. Action at all levels of the systems model is required to ensure this understanding, enforcement, and use. For example, action by educators to ensure journalists and student journalists understand how the existing legislation can be used would sit at the support/peer/community level of the framework. While perpetrators heeding the law and victims using the legislation would be at the individual level of the framework. Education and communication are the key to coordinating responses to enhance their effectiveness. This illustrates how awareness of the interconnectedness allows for a coordinated responses and can enhance the effectiveness of actions at each level. The concluding chapter frames trolling as a ‘wicked problem’ and discusses the importance of communication and education to enhance harmonisation of responses to trolling.
Chapter 5
Conclusion

This study provides a comprehensive analysis of the impact of trolling on journalists in the Australian news media between 2015 and 2021, mapping its emergence and responses to it. The timing of this study was fortuitous, given the rapid growth of technological innovation and subsequent economic restructures of journalism globally over the period of this study. By 2015, trolling had cemented itself as a complex social issue, prevalent in media discourses yet rarely examined in academic research. In a world where journalists are required to have an online presence, a cyber culture had emerged where vicious and frequent trolling is commonplace, and rape threats are regularly received in response to journalistic practice. The literature review established that the cost of cyberhate is more than simple hurt feelings: trolling can result in job loss as well as poor mental health. This study’s findings verify this knowledge. ‘Wicked problems’ such as trolling are complex, dynamic, and difficult to navigate. Thus, any resolution strategies must be addressed in a collaborative approach by multiple key stakeholders.

At its most fundamental level, this thesis concerns the difficulties involved in navigating the journalism profession at a time when online participation is a workplace requirement and yet working in this space often provokes malicious abuse from hostile trolls. In practice, there are problems with reconciling the competing demands faced by social media platform companies and online publishers, as the functionality that allows them to instantaneously publish news to a wide audience, results in them being difficult spaces in which to control targeted abuse. Although current debate continues to discuss the intricate balance between two conflicting principles of freedom of expression and the right to protect oneself from harm, this study dives into these murky waters to uncover the impacts on Australian journalists subjected to trolling as a routine corollary of journalistic practise. The rapidly changing media ecosystem sees abuse
distributed in ways that are difficult for individual users and hosts to monitor and challenging for lawmakers to regulate and enforce.

A grounded theory approach, informed by Ackoff’s (1974) Systems Theory, was used to investigate the impact of trolling on victims and responses by anti-trolling campaigners, Twitter, Facebook, newsrooms, and the legal profession. Ackoff’s theory suggests that social problems cannot be resolved by considering stakeholders in isolation, and that a multi-site and multi-level approach to a problem is more likely to succeed. The research process included interviews with ten trolled journalists, and content analysis of a range of relevant archival evidence, such as news articles, statements by social media platforms, and Australian legislation. Where professional insight into the results of the content analysis was required, the research included interviews with relevant experts.

This thesis aimed to illustrate the very real damage that abusive commentary can inflict on a victim for simply doing their job, and to analyse various response strategies implemented by key stakeholder groups in combatting online abuse. The rich data presented in this thesis has achieved these objectives. The study is divided into seven chapters, framed by an introduction, that explain systems and grounded theory, and this conclusion. First, the literature review provided an historical background of trolling in the Australian news media. This descriptive outline of its progression to the commencement of this thesis in 2015 was the starting point for the reflection and critique in the chapters that follow. The findings were then presented in three parts with the first mapping the emergence of the phenomena in the Australian news media, the second exploring the impacts of trolling through in-depth interviews with ten Australian trolled journalists, and the third investigating responses from key stakeholder groups including anti-trolling campaigns, Twitter, Facebook, Australian newsrooms, and legislation. This final chapter concludes the thesis with a summary of critical findings which provides groundwork future research can extend.

Chapter 4.1 mapped the emergence of trolling in the Australian media industry and from this foundation, considered how the epidemic of online vitriol mirrors a
broader network of fractures within current social, political, and economic ideologies as also argued by Phillips (2015). The chapter examined how the expanded capacity for instantaneous communication has escalated the potential for conflict. Increased interactive functions of Web 2.0 that aimed to promote constructive discourse were hijacked by trolls, who used the new avenues to send targeted abuse to victims. Comment sections on news websites and social media platforms blossomed but quickly became cesspools of hate and bigotry. Trolling flourished because of the known absence of consequences of a behaviour with potentially devastating consequences. Conventional rules of conversation and human interaction are often altered in the online domain, with some users more willing to insult and abuse when masked by anonymity and some happy to abuse under their real identities.

The thematic content analysis of 528 news articles revealed critical time periods in the evolution of trolling which are, in chronological order, the embryonic period, incubation period, manifestation period, social problem period, social epidemic period and saturation period. Findings illustrated that most victims are female and more vulgar insults and aggressive threats are commonly directed at women. Analysis showed that of the nature of the abuse directed at journalists whose gender identities were redacted or omitted from 118 of the articles was consistent with them being female. But without any means of confirming this portion as female, this crucial data may indicate that the true scope of the gender skew described in chapter 4.1 is underplayed. Examination of the evolution of trolling in the Australian news media revealed commonalities with age-old enactments of violence, now played out in the online realm. The trolls’ desire to hurt and disrupt the communication of selected victims to strengthen their sense of power has been compared with the social psychology of conventional schoolyard bullying (Maunder & Crafter, 2018). The sheer scale of online content further exacerbates the trolling problem for which journalists are often outmatched and unprepared.

In chapter 4.2 on the impacts of trolling, the research sought to broaden the understanding of trolling by conceptualising malicious commentary as targeted abuse that can deeply affect victims and manifest into real-life experiences such
as depression and job loss. The research described the severity and vulgarity of comments and threats that are predominately directed towards female journalists and considered how such incidents affect their ability to conduct their work, through thematic analysis of interviews with ten Australian journalists who self-identify as being trolled. Participants revealed that anger and aggression commonly not tolerated or accepted offline began to permeate the online user experience. By documenting tangible impacts including stress, anxiety, and changes to work practices, the research showed that the effects of trolling cannot be reduced to those in the virtual world or regarded as separate from reality, as contemporary communication technologies intertwine both realms. This study argued that new two-way communication models employed by news media organisations are ethically problematic as they enable and often amplify opportunities for unprecedented harm. The occurrence, frequency, commonality, range, and severity of impacts suggest the true scope of the trolling problem is currently not adequately addressed and presents as a worrying concern for career longevity and the welfare of journalists. Through descriptions of personal and professional behavioural impacts, the study revealed that while some journalists ceased writing for a period of time, others became emotionally hardened and often relied on resilience as their primary, or sole, coping strategy. Although technological features play an important role in mediating and shaping trolling abuse and its impact, the persistence of poorly regulated platforms continues to enable the haunting and repeated attacks on victims while they perform basic workplace duties. Such systemic fractures expose a broken system that requires coordinated change from multi-level stakeholder groups within the complex trolling paradigm.

The literature review found unfettered free markets routinely manifest social problems including the gender pay and promotion gap and the rise in casualisation in journalism employment, which particularly affect women. From a feminist perspective, chapter 4.3’s campaign response analysis illustrated that gender-based abuse of women online can be considered an inevitable aspect of an unregulated system. This study described how anonymity has exacerbated the trolling problem through the dehumanisation of internet users, but the underlying
root cause is likely to be imbedded far deeper within social ideologies. In context, it is understandable that the first anti-trolling campaigns that emerged to combat the abuse were feminist, and the study described their evolution in the Australian news media. In accord with feminism’s main intention to stimulate public awareness of gender imbalances that may not yet be recognised or collectively understood by the general population, cumulative examples of journalists speaking out about their personal experiences provided a ‘perfect storm’ for raising awareness about trolling. Considering Foucault’s (1971) theory of the episteme, which describes the significance of optimal timing for a social problem to attain public recognition and media traction, the #mencallmethings campaign went viral on August 30, 2011, and emerged as a pivotal point, shaping public perceptions, and cementing trolling as a recognised social problem, and marking itself as the campaign from which all other anti-trolling campaigns followed. In 2015, there were four racial-based campaigns in a clustered group indicating a continued presence of deep-rooted discrimination in contemporary Australian culture. Government responses were not provoked until “Dolly's Law” campaign on January 3, 2018, after 14-year-old Amy “Dolly” Everett committed suicide after suffering continued online abuse. The chapter explained the limited impact the early isolated campaign had, compared with the success of a number of campaigns that ran together and were collectively more difficult for society to ignore. It is the cumulative effect of social movements that creates public discourse, provokes community debate, incites criticism, and prompts academic inquiry, resulting in widespread social condemnation. Over time the persistence of multi-layered campaigns and activism ultimately influenced the Australian public to re-evaluate its standing on the escalating trolling problem.

With the rise of social media, the distinction between socially accepted behaviours within both private and public spheres has become less clear. Social media companies have been instrumental in dissolving traditional boundaries between the two realms, encouraging users to openly express and share personal thoughts and emotions, with minimal, or sometimes no, mediation. While some newsrooms had expected an online presence to both publish and promote work, this in turn opened the floodgates and exposed journalists to abusive personal attacks for
simply doing their job. Current social media platforms belong to private companies subject to limited government regulation, and technologies that collect and sell user information are increasing. The platforms seem simultaneously overwhelmed by, or rather indifferent to, the rise of violent and hateful content despite frequent announcements about efforts to improve.

Chapter 4.4 on Twitter’s response documents journalists’ widespread frustration with Twitter’s inconsistent enforcement of policies that tolerate and permit abusive content on its service, and its subsequent refusal to publicly explain or account for its action and decisions. The study described Twitter’s twelve responses to trolling on its platform and revealed Twitter was the least-regulated of the two social media platforms investigated, allowing the publication of violent, bullying, demeaning, or otherwise antagonistic material. Prior to March 23, 2015, Twitter relied on a user-monitoring model in which victims were required to report abusive commentary that would then be flagged for external moderation. This protracted self-regulation process facilitated the presence, and amplified the frequency, of trolling on the platform. While the chapter set out the few policy, management and feature responses to trolling by Twitter, it also explained why corporations are often uninterested in making their platforms safer by setting and policing clearer boundaries of acceptable behaviours. The chapter illustrated that Twitter was a direct beneficiary of the circulation of disruptive content through increased engagement on its service, and the data suggested that Twitter may have made a selective decision prior to 2015 to delay its response to trolling. It was not until the negative effects of trolling on Twitter’s popularity and commercial viability became evident on February 2, 2015, that any responsive action was initiated. By examining financial motives, the study highlighted that both the intention and effect of hosting such vitriolic commentary was to attract attention, provoke discussion and increase platform usage. Like Twitter, many social media platforms are often complicit in this process, as contemporary business models depend on a consistent flow of user engagements to entice and retain financial revenue from advertisers. When increased traffic equates to higher profit margins, companies can, and do, face ethical dilemmas, and prioritise trending posts regardless of the offensive or discriminatory nature of their
content. In 2021, extreme trolling content, including rape and violent imagery, are now more readily removed by Twitter’s quality filter and artificial intelligence technology, probably not out of goodwill, but to ensure users continue to engage with and interact on the platform.

Although social media companies can connect marginalised groups in a positive and liberating experience (Day, 2020), the Facebook response chapter described how Facebook’s interactive function was leveraged by trolls and became a fertile environment for malicious and offensive abuse. Facebook was once heralded as the keystone for social connectivity, but failures to eliminate widespread abuse on the platform have seen it become a playground for vitriol and hate-speech, although recent actions have aimed to improve trolling on its platform. The study acknowledged Facebook’s attempts to address multiple stakeholders involved in trolling and cyber-bullying, and grouped responses into five categories: support strategies, feature strategies, functional strategies, research strategies and corporate approach strategies, which were independently described and explored. While proactive moderation strategies using artificial intelligence are developing, effective policing remains difficult, and the onus of moderation predominately falls on users. However, the sheer scale of hate speech on the internet makes it impractical to place the burden solely on victims to report each incident, and more resources could be allocated by social media platforms to better protect their users. Some users have shifted online interactions on the platform exclusively to within Groups or on Pages that are moderated voluntarily by administrators.

In the absence of a consensus or concerted democratic effort to uphold the principles of mutual respect, Facebook risks further becoming an unruly space where trolling thrives. The difficulty for the platform is developing a response to online vitriol that finds a middle ground between ignoring abuse and removing the interactive functions that facilitate it. From a feminist perspective, the study considered Facebook’s misogynistic origins, examining resources delegated to combatting abuse on its platform. Online discrimination, particularly sexism and racism, faced by women and minority groups in contemporary media has a lot to do with the male-dominated corporations within which the harassment takes
place. This persistent power of the patriarchy was illustrated in the #MeToo movement. Ultimately, social media platforms, from their founding objectives to their present-day functions, are a culmination of white, patriarchal, elite practices and the continued dominant presence of abusive commentary suggests that misogyny is still firmly entrenched in the contemporary media space. Although recent media scrutiny has shaped and accelerated a more robust attitude towards the trolling epidemic, there are limits to what can be done by hosting platforms alone. Nonetheless, there is still a scarcity, if not absence, of concrete resolution strategies because it is a systemic problem that needs action on all levels.

Trolling is now a central part of the current Australian news media landscape, with the remaining comment sections on reputable news sites being routinely filtered by journalists and moderators for insults and defamation. In addition, social media accounts now being watched more carefully since the High Court decision in September 2021 that holds news outlets responsible for comments on their Pages (Byrne, 2021). This is a significant workplace problem for Australian journalists.

The study demonstrated that, in 2015, an online culture had emerged where vicious and frequent trolling was commonplace, and where rape threats had become a predictable default response to journalistic practice (Bartlett, 2013; Greig, 2016; Phillips, 2015; Citron, 2014; West, 2015; Williams, 2016; Tuohy, 2019; Hyland, 2021). The newsroom response chapter identified ten newsrooms’ responses to trolling, from basic website feature upgrades to workplace procedural changes that considered the emotional and behavioural impacts of trolling. The ABC was identified a global leader in the hard-line response to trolling and other Australian newsrooms were much slower to respond, opting for user-generated content moderation methods for several years until the sporadic introduction of machine-learning software. The #StopTheTrolls campaign launched by The Daily Telegraph on September 11, 2012, was identified as a significant newsroom response. The campaign received widespread public praise, international media traction and global recognition, and emerged as the catalyst for a shift in public perception of trolling’s prevalence from isolated anecdotal evidence to a recognised social epidemic that required urgent and extensive responses from multiple stakeholder groups. The chapter further unpacked the

Chapter 5 Conclusions
round-the-clock and resource intensive process of comment moderation and highlighted the Australian newsroom *Crikey*, which only permits comments during business hours when a comment moderator is on duty. Although the initial aims of comment sections were to facilitate respectful and informative discussions by deleting insults, defamation and hate speech, the expense of manual moderation has made the process more costly, and even large teams found them difficult to police at scale.

While interactive two-way communication functions enable and drive revenue streams, academic discussions about their sustainability and ethics prompted newsrooms to reassess their value. In doing so several publications adjusted or removed their comment features. These included at least 22 news media outlets. However, the continued presence of comment streams on some publications illustrates varied opinions in the current media industry. Debate still ensues about whether the positives of free speech and reader engagement outweigh the risk of the trolls' vitriol. Some newsrooms opted to handball comment moderation and monitoring to social media companies to deal with, and bear legal responsibility for, but a recent High Court decision has closed this loophole (Karp, 2021, September) and placed responsibility back in the newsrooms' remit. The chapter described how deflecting responsibility to social media organisations did not address the trolls' underlying motives or acknowledge the true scale of online abuse in contemporary journalistic practice. More accurately and more likely, the underlying problem with trolling in newsrooms emerged from an accumulation of resource concerns, including strained staffing, insufficient finances, and inadequate technical aptitude that in unison, newsrooms could no longer cope with. Despite suggestions from eleven newsrooms that social media was a more appropriate environment for audience conversation (*ABC*, 2012; *ABC*, 2017; *Popular Science*, 2013; *CNN*, 2014; *Reuters*, 2014; *Mic*, 2014; *Recode*, 2014; *The Week*, 2014; *The Verge*, 2015; *The Daily Dot*, 2015; *NPR*, 2016; *Al Jazeera English*, 2017), Facebook and Twitter were not exemplary regulators at the time and the chapter concluded newsrooms simply could not cope with the onslaught of vexing commenters through internal means. Although multiple strategies have been adopted to varying degrees, the continued presence and virulence of trolling in
the current media landscape suggests these approaches have not yet been successful in reducing abuse to a manageable level and do not provide any protection for freelance journalists.

The legal response chapter analysed the relevant current Australia legislation and found that while a range of statutes can be drawn upon, use of them (to date) in relation to trolling has been scarce, with experts suggesting that more education of the public and law enforcement officers would enhance legal protection. At present, there is no Australian law, statute or common, specific to trolling; however, criminal laws of sedition, libel and hate speech, and civil laws of harassment, stalking, defamation, discrimination, vilification, blasphemy, and obscenity can be drawn upon for prosecution. Although the current harassment law of "using a carriage service to menace or offend" has been around for 26 years, both the vast number of victims and the sustained and endemic presence of abusive commentary online suggest the current legislation is not an effective deterrent. The research presented the complex, lengthy process that trolled journalists must navigate when commencing judicial action against a troll. In an extension on Pearson and Polden’s (2019) published work, the chapter presented two flowcharts, one criminal and one civil, to distil complex legislative procedures into a practical guide that trolled journalists can refer to when seeking to prosecute their abusers. The study illustrated how arduous legal pathways had proven too daunting for some victims, including high-profile case Zoe Quinn. Analysis of Australian laws that journalists could use when trolled showed existing measures within the Australian legislative system are not yet adequate or effective in preventing online trolling or ensuring sufficient reparation for its victims. Scholarly and societal sources implicitly agree on the harmful effects of trolling (Phillips, 2015; Jane, 2015; Citron, 2014; Barnes, 2018) but fall short on any recommendations for its resolution. Although acknowledgment of trolling as a recognised social problem is a start, there is still a long way to go until lagging legislative reform is updated, implemented, and proven as an effective preventative strategy. The study illustrated that law reform should be considered in integration with social theory models, such as the deterrent effect discussed by David Marr (2015), which suggests perpetrators are more likely to abstain from...
negative behaviour if significant criminal charges are attributed to the act. Trolling remains a legal grey area, and the review of Australian laws that journalists could use when trolled found a disconnect between applicable legislation and the broad public understanding about how such laws are implemented and enforced. The public, and in particular Australian journalists, need adequate policies that recognise the dynamic nature of trolling and its real-life impact. The law is, of course, not the only tool for combatting malicious commentary online; however, it is recognised as a critical stakeholder that is fundamental for future proactive and preventative methods. Online aggression remains a challenging phenomenon for authorities and public policy makers to manage, and layering new legislation announced in 2021 on top of existing laws may be part of the solution, but not be the sole solution. A critical part of the problem is the knowledge of the application of existing laws, and responsiveness of law enforcement officers.

This study concludes that continued discussions, both academic and social, about the meaning, harm, and potential solutions to addressing the trolling problem are critical to raise awareness, educate the public and continue to retain pressure on the key stakeholders to act. As discussed in the literature review and echoed throughout every section of this thesis, education of professionals in all stakeholder groups, such as journalists, lawyers, police, councillors, newsroom managers, can change attitudes and stimulate action. Education can foster synergetic relationships between stakeholder groups that could enable malleable approaches to evolve in response to fluctuating dynamics of the online environment.

This thesis brings together perspectives and contributions that go beyond a purely scholarly interest, including activist and journalistic engagements. The study contributes, in particular, to a consideration of the social impact of trolling and the steps that journalists can take when encountering it as part of their work or personal life. The publication of this data aims to stimulate and facilitate future research on this aspect of online discussions. There is still much work to be done to ensure that journalists, readers, and commenters can have respectful and engaging discussions online without the fear of personal attacks. I hope
journalists, as well as moderators, educators, newsrooms, social media organisations and Australian legislative bodies and media audiences, benefit from this research.
References


Alto, P. (2021, May 16). Feminist activist's request to man to stop smoking
triggers violent reaction in China. *ANI.* https://www.aninews.in/news


References


Ballard, A. (2016, December 8). A journalist was sex-shamed by a troll who shared her nude pictures online, so she took the power back. *MamaMia.* https://www.mamamia.com.au


Bartlett, J. (2013, August 21). Trolls aren’t going anywhere, so we need to learn to live with them. *Wired.* https://www.wired.co.uk


References


Black, P. (2014). Tackling trolls in our imperfect system. *ABC.* https://www.abc.net.au


Bossio, D., & Bebawi, S. (2016). Mapping the emergence of social media in


Brown, A. (2014, August 15). We will not tolerate abuse’: Twitter responds to Zelda Williams’ twitter trolls. *The Daily Express.* https://www.express.co.uk


Bullen, J. (2016, December 14). A motivational speaker who suffers from a rare disease has called out the trolls who used her image in a viral meme. *Evening Standard.* https://www.standard.co.uk


Byrne, E. (2021, September 8). High Court finds media outlets are responsible for Facebook comments in Dylan Voller defamation case. *ABC*. https://www.abc.net.au


Cobus, P. (2020, January 24). Nigerian journalist found hacked to death. *VOA*

Committee to Protect Journalists. (2018, November 1). Digital safety: Protecting against online harassment. CPJ. https://cpj.org


Costello, D. [dickc]. (2015, February 2). We suck at dealing with abuse and trolls on the platform and we’ve sucked at it for years. It’s no secret and the rest of the world talks about it every day. We lose core user after core user by [Tweet]. https://twitter.com

Costello, D. [dickc]. (2015a, February 3). Let me be very very clear about my response here. I take PERSONAL responsibility for our failure to deal with this (trolling) as a company. I thought I did that in my note, so let me reiterate what I said [Tweet]. https://twitter.com


Davidson, H. (2021, May 13). China’s feminists protest against wave of online

https://about.fb.com

Davis, A. (2021, October 13). Advancing our policies on online bullying and harassment. *Facebook.* https://about.fb.com


https://www.schoolgovernance.net.au

https://backlinko.com


https://www.buzzfeed.com/

Diaz, J. (2021, March 15). Thousands march in Australia as another #MeToo wave hits the country. *NPR.* https://www.npr.org/

Diaz, J. (2021, October 14). Facebook will adopt new policies to address harassment targeting public figures. *NPR.* https://www.npr.org/


D’Onfro, J. (2016, August 30). Facebook is telling the world it’s not a media company, but it might be too late. *Business Insider.* https://www.businessinsider.com.au

Dorsey, J. [Jack]. (2018, March 1). We’re committing Twitter to help increase the collective health, openness, and civility of public conversation, and to hold ourselves publicly accountable towards progress [Tweet]. https://twitter.com


Edraki, F., & Carrick, D. (2019, March 22). Trauma of news journalism in focus after The Age found responsible for reporters’ PTSD. ABC. https://www.abc.net.au


Evans, K. (2011, November 11). Men call me things: it’s not as romantic as it sounds. ABC. https://www.abc.net.au


Facebook. (2018c, July 26). Hard questions: Who reviews objectionable content on Facebook – and is the company doing enough to support them? https://about.fb.com/

Facebook. (2019, October 2). Empowering our community to stand up to
bullying. https://www.facebook.com


Facebook. (2021). Why didn’t Facebook remove the content that I reported? https://www.facebook.com


References

Fitzgerald, M. (2022, May 3). Around the world, journalists are resisting the regimes that would jail and kill them. *The Guardian.* https://www.theguardian.com


the hidden decisions that shape social media. Yale University Press.
Green, M. (2018, January 24). No Comment! Why more news sites are dumping their comments sections. KQED. https://kqed.org
Gross, D. (2014, November 21). Online comments are being phased out. CNN.


Hendriks, J. (2021, February 24). Consent isn’t as simple as ‘no means no’. Here’s what you need to know. ABC. https://www.abc.net.au


Higgins, P. (2015, March 13). Hey Twitter, Killing Anonymity’s a Dumb Way to

References

388

https://www.dailytelegraph.com.au

Hobson, J. (2019, February 5). A call for journalists to stop Tweeting. *WBUR.*
https://www.wbur.org


https://doi.org/10.1177/1948550617711227


Hughes, B. (1986). Macroeconomic wage flexibility: comparative international
International Federation of Journalists. (2019, November 22). IFJ launches guidelines to fight back collectively against online trolling of women journalists. IFJ. https://www.ifj.org


Jane, E., & Vincent, N. (2017, July 18). Women online are getting used to cyber hate. They need to get used to reporting it. The Age. https://www.theage.com.au


the bell now trolls. *The Daily Telegraph.*
https://www.dailytelegraph.com.au


Kannan, L. (2021, February 17). After prolonged inaction against abuse, it is time to hold Twitter accountable. *The Pitt News.* https://pittnews.com

Kanter, J. (2019, March 21). Facebook rejects idea of putting time delay on live videos to stop horrors like the New Zealand Mosque shootings from being broadcast to the internet. *Business Insider.*
https://www.businessinsider.com


https://www.theguardian.com

https://www.theguardian.com

Keep, E. (2012, February 17). Troll me once, shame on you, troll me twice. *ABC.*
https://www.abc.net.au


Kirschbaum, E. (2016, September 6). In Germany, it can be a crime to insult


Latimore, J. (2018, August 10). Australia is deplorably racist, as people of colour are reminded when they speak up. The Guardian. https://www.theguardian.com


Lieberman, M. (2020, October 9). A growing group of journalists has cut back on Twitter or abandoned it entirely. Poynter. https://www.poynter.org


References


McGoogan, C. (2017, June 30). Germany to fine Facebook and YouTube 50m if they fail to delete hate speech. *The Telegraph*. https://www.telegraph.co.uk

McGrath, P. (2014, October 23). GamerGate: Female gamers and developers threatened with rape, murder as online row rages. *ABC*. https://www.abc.net.au

McKinnell, J. (2019, June 24). Former detainee Dylan Voller gets court win against media giants over Facebook comments. *ABC*. https://www.abc.net.au


Macmillan, J. (2020, December 23). Australia]’s eSafety commissioner would be able to remove seriously harmful content under proposed cyber abuse laws. *ABC*. https://www.abc.net.au


https://blog.medium.com

https://www.npr.org


https://www.smh.com.au


https://www.theguardian.com


https://doi.org/10.21153/ps2016vol2no1art503


https://twitter.com
https://info.muckrack.com
Mumbrella. (2021, November 12). Mumbrella to reintroduce comments on a 6
Mumby, P. J., (2002). Statistical power of non-parametric tests: a quick guide for
Munn, L. (2020). Angry by design: toxic communication and technical
architectures. *Humanities and Social Sciences Communications.* 7(53). 1-11.
doi:10.1111/2041-9066.12021
Murphy, K. (2016, September 14). Comprehending Pauline is not the challenge.
Engaging constructively with Hansonism is. *The Guardian.*
https://www.theguardian.com
Napoili, P. M. (2019). *Social media and the public interest. Media regulation in the
disinformation age.* Columbian Universtiy Press.
News Corp Australia. (2019, February 7). News.com.au has a fancy, new
News Corp Australia. (2019a, August 24). Pauline Hanson gets stuck climbing
https://www.news24.com
Newton, C. (2014, December 2). Twitter releases improved tools for reporting
abuse and harassment. *The Verge.* https://www.theverge.com
Newton, C. (2015, February 26). Twitter starts tracking phone numbers to
prevent its worst users from creating new accounts. *The Verge.*
https://www.theverge.com
https://www.theverge.com
https://www.theverge.com

References


Nuzzaci, T. (2010). A design thinking approach to leadership development. (Master of Science in Organizational Dynamics Theses). https://repository.upenn.edu/od_theses_msod/41


O’Callaghan, L. (2021, April 16). Australian journalist hits back at trolls after telling Meghan Markle to shut up. *Express.* https://www.express.co.uk

O’Connor, A. (2015, October 6). No, feminists are not ‘pissing their pants for equality’ — it’s a giant hoax. *Yahoo News.* https://uk.news.yahoo.com


O’Reilly, L. (2014, December 11). Facebook: We don’t call them users anymore, we call them people. *Business Insider.* https://www.businessinsider.com


References


Phillips, W. (2015). This is why we can't have nice things: mapping the relationship between online trolling and mainstream culture. MIT Press.


Posetti, J. (2017). Fighting back against prolific online harassment: Maria Ressa. In *An attack on one is an attack on all: successful initiatives to protect journalists and combat impunity*, 37-40. UNESCO.


Powell, A., & White, N. (2020, February 29). Why we’re killing our comments


Rawlins, J. (2011, June 22). Court publishes reasons for releasing internet troll. ABC. https://www.abc.net.au


Ricketson, M., & Wake, A. (2019, March 8). Should news orgs be legally liable for the traumatic situations they put reporters in? A landmark court decision

References
in Australia says yes. *Nieman Lab.* https://www.niemanlab.org


Rigby, B. (2020, June 30). Junkee will no longer publish journalists' bylines on some stories to protect them from harassment campaigns. *Mumbrella.*
https://www.mumbrella.com.au


https://doi.org/10.1145/3274417


https://www.theguardian.com


Roth, A. (2018. August 2). Russian journalists killed in CAR 'were researching military firm'. *The Guardian*. https://www.theguardian.com


Schapals, K. A., & Bruns, A. (2019, October). Are journalism metrics trustworthy enough to be used in editorial decision-making? In *Annual Conference of the*
Association of Internet Researchers. Brisbane, Australia.


Soni, J. (2013, August 26). The reason HuffPost is ending anonymous accounts.
Huffington Post. https://huffpost.com
Srinivasan, R. (2017. February 27). Silicon Valley has designed algorithms to reflect your biases, not disrupt them. Quartz. https://qz.com
Statt, N. (2017, November 7). Facebook’s unorthodox new revenge porn defence is to upload nudes to Facebook. The Verge. https://www.theverge.com
Stevenson, R., & Howcroft, R (2021, September 17). 'Can be a cesspit': Why journalists are quitting Twitter (and why that's a bad thing). 3aw. https://www.3aw.com.au
Stewart, J. (2015, March 10). #IstandwithMariam: Hate groups launch global attack on Muslim advocate after singlet protest. ABC. https://www.abc.net.au

References


Taylor, D., & Mitchell, S. (2019, September 11). 1,400 cases of ‘revenge porn’ abuse reported across Australia. ABC. https://www.abc.net.au

Taylor, J. (2020, June 1). Australian media companies face defamation liability for comments on Facebook after court dismisses appeal. The Guardian.


Teach Us Consent. (2021). Petition for consent to be included in Australian
schools’ sex education earlier. https://www.teachusconsent.com/

References


Twitter Support. (2019, July 24). We're removing Quality Filter from search settings because people told us the filter was confusing and they didn't want it. We've been improving search over time so Top shows you the best Tweets without the extra setting and Latest shows [Tweet]. https://twitter.com


Vernon, J. (2012, February 10). When is a problem 'wicked' as opposed to 'tame'? *The Open University*. https://learn1.open.ac.uk


White, N. (2022, February 8). Social media moderators are like the paramedics of the internet – and they’re right in the firing line of trolls. *ABC*. https://www.abc.net.au


Wondracz, A. (2019, February 9). ’I was mortified’: Travel journalist, 54, reveals her revenge porn horror after ex-boyfriend posts naked photographs of her on a swingers website. *The Daily Mail*. https://www.dailymail.co.uk


Worrall, R. (2020, June 2). Media companies can now be held responsible for your dodgy comments on social media. *The Conversation*. https://www.theconversation.com


Appendix one

INFORMATION LETTER TO PARTICIPANTS

Title of thesis:

*A study of the emergence, impacts and responses to trolling in the Australian news media.*

My name is Delysha Pick and I am a PhD student enrolled in a Doctor of Philosophy at Edith Cowan University in Perth, Western Australia.

You are invited to take part in this research project, which I am conducting as part of the requirements of my doctorate. The research project has ethics approval from the Faculty of Education and Arts Ethics Sub-Committee.

This project aims to study the impacts of trolling on journalists and what is best-practice in Australia.

If you choose to take part in the project you will be asked a series of semi-structured interview questions. The interview will be recorded and transcribed and should take no longer than one hour to complete.

All information collected during the research project will be treated confidentially and will be coded so that you remain anonymous. All data collected will be stored securely on ECU premises for five years after the project has concluded and will then be confidentially destroyed. The information will be presented in a written report, in which your identity will not be revealed. You may be sent a summary of the final report on request.

I do not anticipate any risks associated with participating in this research project. Participation in this project is voluntary and you are free to withdraw at any time and there will be no penalty for doing so.

If you would like to take part in the project, you will first be sent a consent form via email. By reply to the email, and agreeing to an appropriate interview time, this will indicate you have read the consent form and that you give permission for your answers to be analysed in this project.
If you have any questions about the research project or require further information you may contact the following:

Student Researcher: Delysha Pick

Supervisor: Kayt Davies

Supervisor: Trevor Cullen

If you have any concerns or complaints about the research project and wish to talk to an independent person, you may contact:

Research Ethics Officer
Edith Cowan University
270 Joondalup Drive
JOONDALUP WA 6027
Phone: [Redacted]
Email: research.ethics@ecu.edu.au

Thank you for your time,

Yours sincerely,

Delysha Pick
Appendix two

CONSENT FORM

Title of Thesis:

_A study of the emergence, impacts and responses to trolling in the Australian news media._

- I have been provided with a letter explaining the research project and I understand the letter.
- I have been given the opportunity to ask questions and all my questions have been answered satisfactorily.
- I am aware that I can contact Trevor Cullen or Kayt Davies if I have any further queries, or if I have concerns or complaints. I have been given their contact details in the Information Letter.
- I understand that participating in this project will involve answering a set of semi-structured interview questions, that will be recorded and transcribed.
- I understand that the researcher will be able to identify me but that all the information I give will be coded, kept confidential and will be accessed only by the researcher and his/her supervisor.
- I am aware that the information collected during this research will be stored in a locked cabinet at ECU for 5 years after the completion of the project and will be destroyed after that time.
- I understand that I will not be identified in any report, thesis, or presentation of the results of this research.
- I understand that I can withdraw from the research at any time without penalty.
- I freely agree to participate in this project:

Please note by returning the questionnaire online, this will illustrate you have read and agreed to the terms and conditions of this consent form.

No written signature is required.
Appendix three

Semi-structured interview questions addressing research Question 2 -
“*What are the impacts for trolled journalists in the Australian news media?*”

1. How prevalent do you feel is trolling in your workplace?
2. How prevalent do you feel in journalism in general?
3. What are your thoughts on the suggestion trolling is on the rise?
4. Can you suggest any reasons for this?
5. How do you feel trolling has affected you emotionally?
6. How do you feel trolling has affected you at work?
7. How do you feel trolling has affected your writing?
8. At times have you filtered what you write due to this?
9. What are your views on the "don't feed the trolls" advice?
10. What are your views on the “freedom of speech” defence of trolls?
11. Can you give examples of the trolling language or phrases you've experienced?
12. Was there any support offered in the workplace?
13. Was there any support systems in place in the workplace in regards to trolling?
14. What are the workplace procedures in regards to trolling?
15. What advice would you give another person experiencing trolling?
16. What support helped best?
17. Do you know of anyone who has received compensation at present, in your newsroom or other?
18. What are your views on trolling being considered for workplace trauma?
19. Is there anything else you would like to add?