# Rapid Evidence Assessment: The Prevalence and Impact of Online Trolling

## Department for Digital, Culture, Media and Sport

**Reference No: 101030**

Centre for
Strategy & Evaluation
Services

Centre for Strategy & Evaluation Services LLP
Westering House
17 Coombe Road
Otford, Kent TN14 5RJ
United Kingdom
**E:** enquiries@cses.co.uk
**T:** +44 (0) 1959 525122

# Table of Contents

Executive Summary i

1. Introduction 1

1.1 Overview of the report 1

1.2 Objectives and scope of the study 1

1.3 Quality of the evidence base 2

2. Rapid Evidence Assessment 5

2.1 In the literature, what are the different definitions and types of trolling? 5

2.2 What is the prevalence of trolling and does this vary by type of social media platform? 7

2.3 What is the profile of 'typical' trolls 12

2.4 Is trolling a stepping stone/gateway to other negative behaviours? 15

2.5 Can any differences be identified in the online and offline behaviour of trolls? 15

2.6 What is the profile of typical online trolling victims 16

2.7 What impact does trolling have on victims online and offline behaviour? 18

2.8 Can any practical methods be identified to challenge trolling? How effective have past interventions been? 20

3. Conclusions and Recommendations 24

3.1 Overall Conclusions 24

3.2 Key Findings on Research Questions 24

3.3 Recommendations 27

Appendix A: Overview of Literature 29

Appendix B: Definitions of online abuse identified within the literature 49

## Tables

Table 1.1: Search terms for the REA 2

## Boxes

Box: 1.1: Key Research Questions 1

Box: 1.2: Quality of the evidence base 3

Box: 2.1: Definitions of trolling identified within the literature 5

Box: 2.2: The Four Characteristics of Trolling (as defined in Hardaker, 2010) 7

Box: 2.3: Conclusions - What are the different definitions and types of trolling? 7

Box: 2.4: Pew Research Centre evidence on prevalence online trolling 9

Box: 2.5: Varying estimates of prevalence of online trolling in other European countries 11

Box: 2.6: Conclusions - What is the prevalence of trolling and does this vary by type of social media platform? ... 12
Box: 2.7: An explanation of the “Dark Tetrad” ... 13
Box: 2.8: Conclusions - What is the profile of ‘typical’ trolls ... 14
Box: 2.9: Conclusions - Is trolling a stepping stone/gateway to other negative behaviours? ... 15
Box: 2.10: Conclusions - Can any differences be identified in the online and offline behaviour of trolls? ... 16
Box: 2.11: Conclusions - What is the profile of typical online trolling victims ... 18
Box: 2.12: Conclusions - What impact does trolling have on victims online and offline behaviour? ... 20
Box: 2.13: Conclusions - Can any practical methods be identified to challenge trolling? How effective have past interventions been? ... 22
Box: 3.1: Defining Trolling ... 24

# Executive Summary

Provided below is a summary of the findings of a 'Rapid Evidence Assessment: The Prevalence and Impact of Online Trolling'. The assignment was carried out in 2018 for the Department for Digital, Culture, Media and Sport (DCMS) by the Centre for Strategy and Evaluation Services (CSES).

The Rapid Evidence Assessment had two principal objectives, namely to: establish a working definition of trolling, including both hate content and online abuse; and, secondly, to develop a better understanding of the nature, prevalence and impact of online trolling, as well as developing a profile of "typical" trolls and their victims. A number of more specific questions were defined in DCMS's terms of reference and these are summarised below:

### Key Research Questions

- What is the prevalence of trolling and does this vary by type of social media platform?
- What is the profile of 'typical' trolls (may include motivation, rationale for choosing victims, number of victims, prevalence of trolls etc)
- Is trolling a stepping stone/gateway to other negative behaviours?
- Can any differences be identified in the online and offline behaviour of trolls?
- What is the profile of 'typical' victims (may include gender, age, political beliefs, religious beliefs, etc)?
- What impact does trolling have on victims online and offline behaviour?
- Can any practical methods be identified to challenge trolling? How effective have past interventions been?

## 1. Overall Conclusions

**The term trolling is used in the literature to cover a broad range of abusive online activities, with no uniform definition being applied.** The majority of papers examined include some definition of trolling, or at least define a sub-category of trolling, e.g. gender-trolling, flaming, etc. Trolling is a relatively recent phenomenon and one which appears to be evolving, from a slightly anarchic form of activity carried out by individuals towards a more automated form of online abuse. With regard to developing a working definition of trolling, some common parameters emerge from the literature review:

### Defining Trolling

- **Context:** trolling is an activity which is carried out online and is associated with activities where debate is encouraged (e.g. social media platforms, online forums, discussion and comment threads, online gaming chat groups etc).
- **Anonymity**: there is usually a perception of anonymity associated with the perpetrator, or the relationship is distant, for example in the case of attack on disempowered social groups. Counter to this, the victim is vulnerable to exposure as trolling is a public act.
- **Activity:** trolling involves posting off-topic material, inflammatory or confusing messages.
- **Motivation:** trolling can be used to create disruption and discord, to provoke a response from individuals or groups of users, or as a silencing tool to discourage other internet users from getting involved with additional online discussion. Trolling may be undertaken for

amusement or in order to cause harm to specified targets.

**The Rapid Evidence Assessment on trolling underlines a lack of robust, inter-disciplinary study on trolling, in terms of both impact and prevalence.** While the term is often used in academic literature, it lacks clear definition and appears to be used more as a "keyword" to draw attention to study on a plethora of subjects related to online activity than as a developed subject of academic study in its own right.

## 2. Key Findings on Research Questions

### 2.1 What is the prevalence of trolling and does this vary by type of social media platform?

**Estimates of prevalence identified within the literature vary widely, according to the definition of trolling used and the population studied (including factors such as age, gender, physical location and social group).** Prevalence is also measured according to different metrics (number of perpetrators, number of victims and level of activity), which can make it hard to draw meaningful comparisons between studies. The literature demonstrates a clearer focus on cyber-bullying (possibly because this is a better-defined term and therefore easier to measure), with limited attempts to quantify the prevalence of online trolling specifically. Furthermore, the quality of the evidence is not strong.

**A principal methodological weakness shared by many studies providing estimates of the prevalence of online trolling is that they rely on surveys which could be affected by selection biases.** None of the studies were found to provide confidence intervals for their estimates of prevalence, which would allow some understanding of how small or large the actual prevalence might be in the population.

**With regard to the UK, one of the most useful estimates of prevalence of online trolling comes from the 2017 Ofcom report on Adults' Media use and attitudes.** Two data sources have been used to inform the report: a survey of 1,846 adults aged 16 and over, and results from Ofcom's Technology Tracker based on another survey of 3,743 adults aged 16 and over. According to the study, 1% of Internet users in the UK have been trolled online at least once over the past 12 months. This goes up to 5% for respondents aged 16-24.

**Looking beyond studies focusing on the UK, there is more evidence on the extent of online trolling from other countries.** A number of studies could be identified estimating prevalence of online trolling and related behaviour in other European countries. Many more studies were identified estimating prevalence of cyberbullying in other countries than the UK or estimating prevalence without reference to specific geographies. Other studies report incidence rates of 31.4% for cyberbullying and 24.6 to 30.2% for cyber victimisation.

### 2.2 What is the profile of 'typical' trolls

**Establishing the profile of a 'typical' troll is difficult, with at least fourteen different definitions of trolling put forward in the existing literature (excluding activities which incorporate trolling but are broader than this specific subcategory).** Nevertheless, a number of common traits are referenced in many of the papers, largely at the individual level but also some considering broader factors such as socio-economic indicators. Many of these traits are based on qualitative interviews and surveys, which rank character traits demonstrated by trolls against psychological indicators. This type of research should be treated very carefully, as it can be subjective and runs the risk of reflecting the researcher's own prejudices.

A number of indicative psychological traits were consistently referred to in studies - most noteworthy is the 'dark tetrad' of narcissism, psychopathy, Machiavellianism, and everyday sadism.

An alternative view on the profile of trolls was put forward by Cheng et al (2017), who proposed that the majority of them are typical internet users, rather than a subset of atypically antisocial individuals.

**A number of possible motivations for trolling can be found in the literature reviewed for this study. These include**: a need for attention, everyday sadism, low self-confidence, lack of empathy, and a desire for amusement. It is important to note that specific kinds of trolling have different motivations, representing heterogeneity within the trolling community.

## 2.3 Is trolling a stepping stone/gateway to other negative behaviours?

**There is limited literature dealing with trolling as a stepping stone to further negative behaviours, suggesting that it is difficult to assess the direction of association.**

**There is a lack of existing research on whether trolling is a gateway to other negative behaviours (online or offline).** This makes it difficult to assess to what extent there is a clear causal link between trolling and other negative behaviours. That the majority of the texts available are either reviews of secondary literature or based on surveys of self-described trolls is also problematic – the former fails to generate additional data to analyse, and the latter has the risk of trolls responding mischievously to researchers.

## 2.4 Can any differences be identified in the online and offline behaviour of trolls?

**There is evidence suggesting that online and offline identity are closely related and cannot be totally decoupled.** On a related note, several studies found links between online and offline bullying. At the same time, this continuity of behaviour has to be considered alongside the importance of anonymity to trolls. As a result of the 'protection' of the screen, studies show that users' inhibitions are lowered online due to the lower risks of discovery and consequences.

**However, insufficient work has been carried out to link together trolls' presence online and offline.** Furthermore, the majority of papers which consider links between these behaviours tend to focus heavily on qualitative research based on textual analysis or interviews, or deal with secondary literature. While this is useful in providing some detailed insights into how trolling works in individual cases, findings would be further bolstered by robust quantitative research (based on a clearly defined research question, a well thought out and transparent methodology, supported by a large and representative dataset) into online and offline behaviour.

## 2.5 What is the profile of typical online trolling victims

**The literature suggests that certain groups within society are particularly affected by online trolling and the related phenomenon of cyber-bullying.** These groups include women, ethnic minorities, religious groups, people with disabilities, and the LGBTQ community. Survey findings suggest that age (younger) and gender (women) in particular are most closely associated with the experience of online harassment – and with trolling as a subset of online harassment. In terms of gender, several papers were identified suggesting that women are targeted by online trolls and cyber-bullying more often than men.

**Apart from gender and age, another way of developing a profile of typical trolling and cyber-bullying victims is to identify characteristic psychological traits.** One study examined found that cyber victims show higher levels of neuroticism, including high anxiety, hostility, depression and vulnerability to stress, high agreeableness (altruism and modesty), and higher levels of openness – however it is hard to know to what extent victims were already more likely to display these traits and to what extent they might be caused by being the target of trolling. A few studies identified also highlighted marginal groups in society that were more likely to be trolled or cyberbullied.

## 2.6 What impact does trolling have on victims online and offline behaviour?

**There is very little quantitative evidence available regarding the impact of trolling on victims' online and offline behaviour.** Instead, the literature tends to focus on case studies, anecdotal evidence and self-reporting by victims of how they perceive trolling to have affected them.

**There does appear to be some evidence of continuity between online and offline behaviour, both for trolls and victims of trolling.** For trolls, the literature suggests this lies in a correlation between their tendency to be abusive online and offline. Specifically, there appears to be evidence of a spill-over between online abuse in some contexts and offline abuse such as hate mail, death threats and stalking. The literature on psychological traits of trolls suggests that those underlying tendencies elicit problematic behaviour both online and offline, and that is why this correlation appears but little direct evidence is available to support this hypothesis as yet. For victims, trolling may lead to a reduction in engagement online and social withdrawal offline. In situations where online and offline abuse are happening in combination, the impact on the victim tends to be more significant. This may be linked to the fact that online activity affects areas of an individual's life which are usually perceived as private, meaning that online bullying may be more pervasive. Voggesser *et al* 2017 found that the online aspect of abuse nullified attempts to escape from offline bullying.

**A complication in seeking to assess impacts is that in much of the literature, no distinction is made between trolling and cyber-bullying – terms which are sometimes used interchangeably or used separately to refer to similar abusive online behaviours.** Indeed, online trolling appears to engender similar reactions in its victims to offline harassment.

**Commonly reported symptoms amongst trolling victims include increased emotional distress and embarrassment, and increased risk of clinical or subclinical symptoms (such as those for depression, anxiety and posttraumatic stress disorder).** A number of the studies identify similar behavioural impacts, including substance abuse, shame, humiliation, low self-esteem, paranoia, withdrawal from social life and detrimental impacts on personal relationships (for example with other family members, intimate partners).

**Even when victims show relatively high levels of psychological resilience, they may still change their online behaviours in response to trolling**. Furthermore, even those who are resilient to trolling themselves may change their behaviours in response to threats directed at others (for example, family members or those within their communities).

## 2.7 Can any practical methods be identified to challenge trolling? How effective have past interventions been?

**The literature reviewed highlights two types of approaches to combat online trolling.** The first relies on automatic mechanisms that rely on algorithms or other forms of artificial intelligence, whereas the second involves human-based interventions or 'soft' measures that require deliberate action by individuals in response to specific occurrences, such as online-community group decisions to ignore trolls.

**Many studies look at legal protection for victims of trolling in different countries, as well as how this has been implemented in practice.** There is a consensus among legal scholars that the victim is the one who needs to take responsibility to deal with harassments and threats online. In February 2018 the UK government asked the Law Commission to review the laws around offensive communications and assess whether they provide the right protection to victims online.[1]

[1] For more information on this, see the Prime Minister's speech on the matter in February 2018 or the press release published by the Law Commission

**Very few papers were found which assessed effectiveness or efficiency of reporting mechanisms on social media platforms, suggesting there is little-to-no research on how effective this is.** An exception to this was one study on Wikipedia which tested the effectiveness of a suggested algorithm against a manual sample in identifying "troll posts". Another paper on social bots on Twitter tried to measure the percentage of success of reporting hybrid social bots.

**Furthermore, there is limited information available on how major online platforms address trolling and the effectiveness of the strategies currently being used.** A review of transparency reports published by Facebook, Twitter and Google provided little clear insight into trolling activities on this platform and instead reported requests filed by the UK government for criminal investigation, shutting down accounts etc. This lack of publicly available information may have affected the methods scholars use to assess prevalence of trolling and effectiveness of strategies to counter it.

**Attempts to combat trolling identified in the research have tended to look at a combination of top-down approaches such as vetting of comments before publication or the imposition of codes, or bottom-up approaches where users work together to neutralise the activities of trolls.** There is also an increased interest in the use of automated IT tools to identify and neutralise trolls, although these have yet to be tested in practice. The legal position of the UK government regarding offensive speech online is the subject of a recently published scoping report by the Law Commission, which found that most online abuse is covered by existing law relating to malicious communication but that practical and cultural barriers often prevent harmful online content from being recognised, pursued and prosecuted to the same extent as offline abuse.[2]

**Given the gaps in what is known, more research is needed to gain a clearer understanding of what trolling is, how it is affecting UK society and what needs to be done to effectively counter its negative impacts.** Nonetheless, the literature reviewed as part of this research has highlighted some common findings and areas of agreement amongst researchers working in these areas, as well as pointing to some interesting hypotheses which could be explored in further study.

The quality of evidence is uneven, reflecting the different kinds of data needed to reach conclusions for different aspects of trolling. One of the strongest aspects is the study of the profiles of typical trolls. Ready quantitative and qualitative data is available through surveys and interviews with trolls. Others, such as establishing whether online trolling leads to negative offline behaviours, are far harder to establish: correlation does not equal causation, and the negative offline behaviours and online trolling may both be the result of a third factor. Complicating this further is the heavy prevalence of papers which simply analyse and synthesise earlier studies rather than capturing new data. Replication of previous studies is important in order to further establish different aspects of trolling.

## 3. Recommendations

Based on our review, some suggestions are made below with regard to future research and policy actions on this topic:

- A UK-wide survey is needed to better understand the prevalence and impact of trolling, based on a clear definition of trolling.
- More research into cross-platform comparisons of trolling activities and attempts to counter them would provide valuable insights into common strategies and differences across platforms, and what prevention measures might be transferable from one platform to another.

[2] Law Commission (UK), Abusive and Offensive Online Communications: A Scoping Report, November 2018, available at: https://s3-eu-west-2.amazonaws.com/lawcom-prod-storage-11jsxou24uy7q/uploads/2018/10/6_5039_LC_Online_Comms_Report_FINAL_291018_WEB.pdf

- The literature is fragmented between different disciplines, with little evidence of inter-disciplinary research. An inter-disciplinary approach covering legal, psychological, technological and other aspects of trolling would be welcome.
- There is some evidence of a link between online trolling behaviours and traditional bullying. More research into this aspect should be considered as it would inform future prevention strategies (i.e. whether "soft" preventative approaches such as the provision of psychological support should be provided or whether "hard" technological solutions are better suited to dealing with the problem).

# 1. Introduction

## 1.1 Overview of the report

**This document contains the final report on the assignment 'Rapid Evidence Assessment: The Prevalence and Impact of Online Trolling'. The assignment was carried out in early 2018 for the Department for Digital, Culture, Media and Sport (DCMS) by the Centre for Strategy and Evaluation Services (CSES).**

The report is structured as follows:

- **Section 1: Introduction** – the rest of this section provides a summary of the study objectives and scope and the research undertaken by us.
- **Section 2: Rapid Evidence Assessment** – this section provides an assessment of the evidence currently available on the prevalence and impact of online trolling. The assessment is structured around the seven research questions set out in DCMS's terms of reference.
- **Section 3: Overall Conclusions** – the final section provides a summary of the key findings together with conclusion and recommendations.

The report is supported by an appendix setting out a summary of each piece of literature that we identified as falling within the scope of the research. A total of 69 literature sources are listed.

## 1.2 Objectives and scope of the study

The Rapid Evidence Assessment (REA) had two principal objectives, namely to:

- Establish a working definition of trolling, including both hate content and online abuse;
- Develop a better understanding of the nature, prevalence and impact of online trolling, as well as developing a profile of "typical" trolls and their victims.

A number of more specific questions were defined in DCMS's terms of reference for the REA and these are summarised below:

**Box: 1.1: Key Research Questions**

- What is the prevalence of trolling and does this vary by type of social media platform?
- What is the profile of 'typical' trolls (may include motivation, rationale for choosing victims, number of victims, prevalence of trolls etc)
- Is trolling a stepping stone/gateway to other negative behaviours?
- Can any differences be identified in the online and offline behaviour of trolls?
- What is the profile of 'typical' victims (may include gender, age, political beliefs, religious beliefs, etc)?
- What impact does trolling have on victims online and offline behaviour?
- Can any practical methods be identified to challenge trolling? How effective have past interventions been?

In terms of scope, the assignment was to examine the difference between online and offline behaviour of trolls, as well as comparing trolling to more traditional forms of bullying, abuse and hate crime. A number of more specific parameters were defined relating to the search terms for the REA:

**Table 1.1: Search terms for the REA**

| | | |
|---|---|---|
| • Troll*<br>• Trolling<br>• Internet trolling<br>• Cyber-trolls | • Adult antisocial Internet behaviour<br>• Online abuse<br>• Computer-mediated communication | • CMC<br>• Cyberhate<br>• Cyber-bullying + trolling<br>• Bots<br>• Adult + trolling + trends |

The terms above were used in combination with terms to draw out evidence such as 'evidence', 'research', 'systematic review', 'rapid evidence assessment', and effectiveness such as 'evaluation', 'assessment', 'impact', 'intervention', 'motivation' as well as ‘prevalence’ and ‘impact’.

With regard to the inclusion/exclusion criteria for the REA, the research was limited to literature written in English in order to better focus on the UK. A five year cut-off limit was also imposed with regard to publication dates, although one or two earlier papers which were consistently referenced in the literature reviewed have been included in this report where relevant (for example, Hardaker 2010) Whilst the primary focus of the literature search was papers considering trolling in the UK, the limited nature of this research and the international nature of trolling as a phenomenon meant that papers from other OECD countries were also examined. The search focused on literature related to adults over the age of 16, and literature published more than five years prior to the start of the study was not examined.

## 1.3 Quality of the evidence base

In total, some 69 papers and studies were identified that fitted the inclusion/exclusion criteria explained in Section 1.2 from more than 1,000 papers identified by using the keyword searches above.

The documents were further reviewed according to their relevance, the methods used for data collection and analysis, and the quality of argumentation put forward. In the first instance, priority was given to finding studies from which trends could be identified. However, due to a dearth of such research in the literature covered, a secondary search was undertaken to identify more theoretical papers or those based on smaller samples and involving qualitative data collection methods (interviews, qualitative surveys, etc). These papers provide a useful starting point for observations and the development of hypotheses, which can then be tested in larger studies. However, they are limited in the scope of their application. For this reason, one of the main recommendations of this REA is the need for a UK-wide survey to better understand the prevalence and impact of trolling, based on a clear definition of trolling.

When analysing the papers which match the proposed inclusion/exclusion criteria, consideration was given to elements such as the methodology used (with reference to a hierarchy of evidence, as exemplified by Bagshaw and Bellomo/Pettigrew and Roberts)[3], sample size and sampling method, use of empirical data or theoretical argumentation, due consideration of alternative arguments and

[3] Bagshaw and Bellomo 2008, p 2 and Petticrew and Roberts 2003, p 527 via Sandra Nutley, Alison Powell and Huw Davies, University of St Andrews (2012) What counts as good evidence?

counterfactuals, and the nature of the journal in which the paper appeared.[4] This allowed for a more holistic assessment of the quality of evidence available and the exclusion of evidence which appeared to be based on subjective assertion and was not clearly backed up by some form of empirical evidence.

The material included consists largely of academic research papers complemented by some policy documents prepared by the UK government, OECD, the European Parliament and European Commission. Following an in-depth review of paper quality and subject matter, 10 papers were excluded leaving a total of 59 papers to be included in the REA.

Overall, the evidence base is quite disparate and lacks a strong focus or interdisciplinary approach. This may be because trolling is a relatively new field of academic study and the phenomenon itself lacks clear definition and is still in a state of evolution. There is also quite a considerable variation in both the quantity and quality of literature available, depending on the individual research question. A summary of the quality of the evidence base is provided below:

**Box: 1.2: Quality of the evidence base**

- There is a considerable amount of literature available on trolling, and although a lot of it is not specific to the UK, the findings are nevertheless potentially applicable as trolling is an activity which is not limited by national borders;
- The literature found in the search shows a lack of consensus regarding the definition of trolling, or regarding methods for defining “prevalence” and “impact” with regard to trolling;
- In general, the literature focused specifically on trolling is often based on qualitative data, including surveys, reviews of existing literature, interviews and case studies.
- Quantitative data on trolling is limited and tends to focus on small groups with specific identities (e.g. gamers, feminists, Muslims, members of a certain online community)
- More robust quantitative surveys (based on large and representative datasets, well-targeted questions, and appropriate and transparent sampling methods) tend to include questions on trolling as one of multiple questions on a much broader topic (for example, attitudes to online behaviour more broadly).
- There is a more extensive body of evidence on cyberbullying and online abuse than specifically considering trolling. Distinctions between different forms of online abuse are often unclear and where this appears to deal with trolling, this evidence has been included in our findings.
- In much of the literature around trolling and online abuse, there is no clear delineation between adults and children. Literature focused solely on children below the age of 16 has been excluded from this assessment, but where the subjects of a study include those a mixture of children, adolescents, and/or adults this has been included in our assessment (as appropriate).
- The majority of papers identified in the literature search provide insights that may help

[4] In general, hierarchies of evidence tend to rank studies according to the methodology used, giving preference to randomised experiments with clearly defined controls (RCTs) are placed at or near the top of the hierarchy over more qualitative forms of evidence gathering such as case study reports. These hierarchies can be useful for determining “what works”, when an issue is clearly defined and well organised interventions to achieve specific goals have been established. For the purposes of this study, however, they have a limited utility – this is because trolling as a phenomenon lacks clear definition, and attempts to counter it have been sporadic. There is a lack of robust quantitative data on the subject, with most of the literature available falling towards the bottom of a standard hierarchy of evidence. Furthermore, when considering strength of evidence, it is important to avoid the risk of generalising from too small an evidence base. Even a high quality RCT cannot form an effective basis for clear conclusions. Until a stronger body of evidence is available, any judgments regarding trolling must be read in the context of a very limited evidence base.

answering research questions 2, 5 and 7 (see earlier table for key);

- Very little research has been discovered looking into the issues raised by research questions 1, 3, 4 and 6;
- Papers examining research question 7 (measures to prevent trolling) appear to focus on “bottom-up” community management strategies rather than “top-down” strategies either at government level or on behalf of social media companies and corporations. As such, papers tend to focus on prevention through technical solutions and online community-based methods as a means to challenge trolls;
- No studies were found which provide hard data on the prevalence of online trolling (question 1), although some try to estimate this based on surveys or try to define what determines prevalence. To the extent that quantitative estimates exist, they vary and are not directly comparable in many cases because of different sampling methods, different definitions and/or differences in target population characteristics.
- Similarly, research into the impact of trolling (question 6) tends to focus on small-scale studies looking at individual platforms or online communities.

# 2. Rapid Evidence Assessment

**This section presents the results of the REA. The assessment is structured around the seven research questions, drawing out conclusions based on the rapid evidence assessment. The report begins with an exploration of the different definitions of trolling found in the literature.**

## 2.1 In the literature, what are the different definitions and types of trolling?

**The term trolling is used in the literature to cover a broad range of abusive online activities, with no uniform definition being applied.** The majority of papers examined include some definition of trolling, or at least define a sub-category of trolling, e.g. gender-trolling, flaming, etc.

**Several papers focus on the broader definition of cyber-bullying rather than just online trolling.** Here, the definition appears to be more concrete. Scholars agree that cyber-bullying is an aggressive act or behaviour taking place online; involving a repetitive behaviour medium to long-term; to some extent harming the victim; and to some extent the victim knows the perpetrator. Others highlight that cyber-bullying usually involved young people in an educational context (Grimme et. al, 2017).

**Many of the activities referred to as trolling in some papers are also described as cyber-bullying, cyber violence or online abuse in other papers.** This suggests that the terms are to some extent being used in an interchangeable way, depending on each researcher's own interpretations or opinions. It has been suggested that cyber-bullying and trolling are essentially the same phenomenon or that trolling can be used as an umbrella term to refer to various forms of online harassment which aim to create disruption and conflict (Seigfried-Spellar, 2017). Nonetheless, defining characteristics often associated with trolling include a perception of anonymity amongst perpetrators and the intention of disrupting online discussion (Hardaker, 2015; Jonason et. al, 2014; Seigfried-Spellar, 2017; Sest, 2017).

A summary of the various definitions is provided below. It should be noted that this box only contains definitions of "trolling". A more complete list, including broader definitions of cyberbullying and online abuse which encompass trolling behaviours and "sub-types" of trolling can be found in the appendices.

**Box: 2.1: Definitions of trolling identified within the literature**

- To post deliberately inflammatory articles on a social media forum (Maltby, 2016);
- Deliberately attacking others online, typically for amusement's sake (Hardaker, 2015);
- The online posting of deliberately inflammatory or off-topic material with the aim of provoking textual responses and/or emotional reaction (Jane, 2015);
- An attempt to argue with and upset people by posting inflammatory and malicious messages (Maltby, 2017);
- The act of deliberately posting inflammatory or confusing messages on the Internet in order to provoke a vehement response from a group of users' (Shaw, 2013);
- The posting of hateful comments by a person or group along with the more aggressive, premeditated and prepared hate movements undertaken by groups of people (Bratu, 2017);
- A range of antisocial online behaviours that aim at disrupting the normal operation of online social networks and media (Tsantarliotis et. Al; 2016);
- The targeting of defamatory and antagonistic messages towards users of social media (Williams

and Pearson, 2016);

- A distinct new form of antisocial behaviour online (March, 2017);
- Flaming, griefing, swearing, or personal attacks, including behavior outside the acceptable bounds defined by several community guidelines for discussion forums (Cheng et. al, 2017);
- Online harassment against strangers, with the intentions of causing disruption and conflict for entertainment (Seigfried-Spellar, 2017);
- A form of behaviour through which a participant in a discussion forum deliberately attempts to provoke other participants into angry reactions, thus disrupting communication on the forum and potentially steering it away from its original topic (Hopkinson, 2013);
- Practice of behaving in a deceptive, destructive, or disruptive manner in a social setting on the Internet with no apparent instrumental purpose (Jonason et. al, 2014);
- Specific example of deviant and antisocial online behavior in which the deviant user acts provocatively and outside of normative expectations within a particular community; trolls seek to elicit responses from the community and act repeatedly and intentionally to cause disruption or trigger conflict among community members (Fichman and San-Filippo, 2015)

Although the specific definitions applied to trolling may vary, there does seem to be some agreement on specific conditions or characteristics that reflect trolling. These are described below.

In relation to the **conditions/characteristics of trolling**, there appears to be a consensus that: trolling is an activity that takes place online; there is also an association with platforms which involve a discursive element, such as social media and online discussion forums (Bratu, 2017; Cheng et. al, 2017; Hopkinson, 2013), and with gaming (Cook et. al, 2017); and trolling is associated with a perception of perpetrator anonymity (Hardaker, 2015; Jonason et. al, 2014; Seigfried-Spellar, 2017; Sest, 2017). Moreover, trolling involves posting off-topic material, inflammatory or confusing messages (Jane, 2015; Maltby, 2015; Williams and Pearson, 2016).

With regard to the **motivation for online trolling**, for some, trolling is undertaken "as an end in itself", for amusement's sake or as an attention-seeking behaviour (Cook et. al, 2017; Hardaker, 2015; Zezulka, 2016); for others, trolls seek to harm others (Bratu, 2017; Cheng et. al, 2017), and may even be part of a broader and more coordinated hate movement (Bratu, 2017). Trolling can also be used to create disruption and discord, to provoke a response from individuals or groups of users (Cook et. al, 2017; Mantilla, 2013; Tsarliotis, 2016), or as a silencing tool to discourage other internet users from getting involved with additional online discussion (Bratu, 2017).

Claire Hardaker's four characteristics of trolling, as defined in her 2010 study *Trolling in asynchronous computer-mediated communication,* have been used as a point of reference for a number of other academics (Craker and March, 2016; Hopkinson, 2013; March, 2015; Sest, 2017; Voggeser et. al, 2017) when attempting to provide a definition of trolling, suggesting that they may be a useful point of consideration. These are defined in the box below.

**Box: 2.2: The Four Characteristics of Trolling (as defined in Hardaker, 2010)**

- **Deception** - acting differently on an online profile to how one would offline
- **Aggression** - aggressive, malicious behaviour undertaken with the aim of annoying or goading others into retaliating
- **Disruption** - seemingly pointless behaviour which appears to seek to gain attention rather than advance a conversation
- **Success** – achieved by provoking a response to the trolling behaviour. [5] Failure to provoke a response will lead to a troll upping their attempts, until success is achieved.

An emerging trend within the literature appears to be the shift from trolling by individual internet users as defined above towards an automation of trolling activities, through the use of bots. These are defined generally as either simple algorithmic programs or cyborgs, that provide technological assistance for trolls to multiply their trolling efforts by scaling up responses/posting/re-tweeting etc.

Two specific bot types identified in the literature are social bots, which mimic human social media activity, and political bots, which aim to manipulate public opinion by spreading political content or by participating in political discussions online.

**Box: 2.3: Conclusions - What are the different definitions and types of trolling?**

- Cyber-bullying is often used interchangeably with trolling and other descriptors such as cyber-violence or online abuse;
- Diverging connotations of trolling make it difficult to achieve one clear definition. Nonetheless, some recurring characteristics of trolling are emerging. These include: context (online), activity (off-topic, aggressive or inflammatory) and motivation (provoking a response from users);
- One important difference between cyber-bullying and trolling is the concept of anonymity;
- Trolling activities can be amplified by the use of "bots", usually in the form of algorithmic programmes or cyborgs which automatically retweet, re-post etc.

## 2.2 What is the prevalence of trolling and does this vary by type of social media platform?

**It has proved difficult to reach definitive conclusions regarding the prevalence of online trolling.** This is partly due to the lack of a defined time window (e.g. period prevalence, or point prevalence) and population context, which would be needed to provide the necessary parameters for measures of prevalence. This has led to estimates which vary widely from each other or provide such a broad incidence window that little can be drawn from them in terms of useful conclusions. Furthermore, the number of studies looking at trolling is very limited. A larger number of studies provide estimates of the prevalence of cyber-bullying and related phenomena such as cyber hate, which encompass trolling activities, than of trolling per se.

**Estimates of prevalence identified within the literature vary widely, according to the definition of trolling used and the population studied (including factors such as age, gender, physical location and social group).** Prevalence is also measured according to different metrics (number of perpetrators, number of victims and level of activity), which can make it hard to draw meaningful comparisons between studies. The literature demonstrates a clearer focus on cyber-bullying

[5] Perhaps best described by the commonly used phrase "successful troll is successful"

(possibly because this is a better-defined term and therefore easier to measure), with limited attempts to quantify the prevalence of online trolling specifically. Furthermore, the quality of the evidence is not strong. Many estimates are based on unrepresentative samples of certain groups within society (e.g. people with mental disabilities, people in a certain age group, etc.) or samples whose make-up is not clearly specified. Moreover, estimates of prevalence are often used within the literature as a method of justifying the author's own research into a specific aspect of trolling, meaning that more papers uncritically refer to estimates made by other papers than actually trying to estimate prevalence themselves. Figures quoted are often taken at face value to demonstrate the importance of the study subject, without any critical assessment of the robustness of the evidence of prevalence being presented.

**A principal methodological weakness shared by many studies providing estimates of the prevalence of online trolling is that they rely on surveys which could be affected by selection biases.** None of the studies were found to provide confidence intervals for their estimates of prevalence, which would allow some understanding of how small or large the actual prevalence might be in the population. Also, only a few studies provided estimates of online trolling or cyber-bullying prevalence in the UK. More studies found estimated prevalence in other countries or in general, but caution should be applied in inferring specific prevalence rates for the UK based on these surveys, given that prevalence rates in other countries may reflect specific characteristics of these countries that do not apply to the UK.

**Nonetheless, interest in prevalence of online trolling or automated so-called 'social bot activity' in social media has increased since 2010, particularly with regard to political manipulation and propaganda (Grimme et. Al, 2017).** First-hand accounts and media reports of issues such as 'Twitter trolling' provide ample evidence that, in recent years, e-bile (email sent for the purpose of generating anger, vitriol and complaint) has become far more 'prevalent, rhetorically noxious, and gendered in nature' (Jane, 2015). This reflects a general increase in interest on the subject of trolling.

**With regard to the UK, one of the most useful estimates of prevalence of online trolling comes from the 2017 Ofcom report on Adults' Media use and attitudes.** Two data sources have been used to inform the report: a survey of 1,846 adults aged 16 and over, and results from Ofcom's Technology Tracker based on another survey of 3,743 adults aged 16 and over.[6] According to the study, 1% of Internet users in the UK have been trolled online at least once over the past 12 months. This goes up to 5% for respondents aged 16-24. These figures are based on 1,553 responses of individuals who go online (Ofcom, 2017). This estimate may be slightly lower than the actual incidence rate, as the question covering trolling simply asked users whether or not they had personally been trolled, without any clear definition or description of the term. Reporting would therefore depend on a respondent's own definition of what "trolling" is, something which varies widely even between academics studying the subject (as has been elucidated in Section 2.1). It is possible that some respondents may not have understood this term, while others may apply a

[6] This is a government-mandated report, carried out every year since 2003, which tracks changing attitudes to the media amongst adults in the UK. It draws on data from two national surveys: the annual Adults' Media Literacy Tracker survey is the primary source, which is based on computer-aided personal interviews with 1,846 adults aged 16 and over in November and December 2016. This information is supplemented with data from the Technology Tracker, a quantitative survey based on 3743 interviews with adults aged 16 and over in January and February 2017 (of whom 3221 were internet users). As a result, there are some differences in the year-on-year comparisons contained in this report – these are flagged as and when they appear – both in the narrative and in the source notes under any relevant chart. Significance testing at the 95% confidence level was carried out, and any findings detailed were found to be significant to a 95% confidence level. Between 2016 and 2017 the Technology Tracker changed from a survey administered using a pencil and paper approach (PAPI) to one administered through a tablet (CAPI). As this could affect the results shown, any difference between 2016 and 2017 was tested at the 99% level, meaning that there is only a 1% probability that the difference is by chance. In addition to reporting on differences over time, the report also looks at adults in different age groups and socio-economic groups, and compares these to adults overall in 2016, to see if there are any significant differences within these sub-groups. Differences between men and women are also considered.

narrow definition of "trolling" or not view themselves as the direct victim of a trolling incident in which they were implicated. No information was requested from respondents regarding the platforms on which such trolling was taking place.

**Specific social groups may be more frequently targeted by trolls, reflecting wider societal prejudices**. Bratu (2013) suggests that social norms regarding good manners and a willingness to follow instructions, as well as both parties' impressions of the others perceived "power" or status within society may affect motives to troll, trolling conducts, perception of conducts such as trolling, and the impact of trolling on online groups. In itself, this argument is not very compelling as the paper lacks clear evidence to support its claims. Nonetheless, the question of power asymmetry does appear to be reflected in other studies which look at the targets of online trolling activities. Often, these seem to be disempowered groups within a particular society.

One example of this in the UK is the Muslim community. Barlow and Awan (2016) quotes cases of anti-Islamic abuse recorded by the UK-based organisation TELL MAMA (Measuring AntiMuslim Attacks). Of 734 reported cases of anti-Islamic abuse that took place in the UK between May 2013 and February 2014, 599 took place online (Barlow and Awan, 2016). Women and the LGBTQ community also suffer from a higher prevalence of attacks. One study of female students found that 37.8% of them had experienced cyber-bullying, and 56% had witnessed it as bystanders (Pilkey, 2012).

### Box: 2.4: Pew Research Centre evidence on prevalence online trolling

According to research by the Pew Research Centre in the USA (Duggan 2014, sample of 3,217 respondents),[7] 3% of adult internet users have seen someone be harassed in some way online and 40% have personally experienced it. Pew Research asked respondents about six different forms of online harassment.

Those who witnessed (experienced) harassment said they had seen at least one of the following occur to others online:

- 60% of internet users said they had witnessed someone being called offensive names (27% experienced this themselves)
- 53% had seen efforts to purposefully embarrass someone (22% experienced this themselves)
- 25% had seen someone being physically threatened (8% experienced this themselves)
- 24% witnessed someone being harassed for a sustained period of time (7% experienced this themselves)

The research also suggests that online harassment occurs particularly frequently on social networking sites or apps (in case of 66% of internet users who had experienced harassment online) and in the comments section of a website (22%).

Meanwhile, in the 2017 report (n = 4,248), 41% of U.S. adults report having personally experienced online harassment—a 1% increase since 2014. 66% have witnessed these behaviours directed at others. 13% report ceasing their social media use after witnessing such events

[7] This report is based on data from the Pew Research Centre's American Trends Panel, a probability-based, nationally representative panel. The survey was self-administered online, with a margin of error of plus or minus 2.4 percentage points.

**As regards political trolling, only two studies were identified that investigated this phenomenon.** A mapping exercise of 28 countries carried out by the Oxford Internet Institute in 2017 (Howard and Bradshaw, 2017) found that all of them had some form of government-sponsored online propaganda programme, including trolling activities. However, no clear estimate of prevalence was possible – partly due to a lack of quantifiable evidence on the subject. Llewelyn et. al (2018) tried to estimate prevalence of Tweets by Russian-controlled Twitter accounts in wake of the Brexit referendum. This study takes a different approach to measuring prevalence than those outlined above, using the number of tweets, rather than number of perpetrators or number of victims, as its main indicator. The study found Brexit-related content on 419 of 2,752 Russian-controlled accounts[8], leading to 3,485 identified tweets gathered between the 29th August 2015 and 3rd October 2017, or 0.005% of all Brexit-related tweets in this period.[9] While the study does not say what the impact of this trolling was, it does point to the way that social bots can produce a large number of messages in a short timeframe and in a specific context. The ability to draw broader conclusions on the impact and prevalence of political trolling, however, are limited due to lack of evidence.

**Having looked at the studies which focus specifically on trolling, we now turn to estimates of prevalence in the broader literature related to cyber -bullying and related phenomena such as cyber hate.** Cyber-bullying prevalence rates in various sample populations based on opinion surveys vary quite substantially from single digit percentages to more than 30% depending on the type of cyber-bullying, the type of sample chosen, and other factors. One report looking at cyber-bullying in the UK identified rates of cyber-bullying victimisation of between 15 and 28% (Williams and Pearson, 2016). In comparison, a systematic review of international research into cyber-bullying found that prevalence rates of victimisation ranged from 4-72%; this broad range points to the need for further review of how trolling is operationalised and measured. The author of the study also quotes one other piece of research which showed that 67% of 15-18-year olds had been exposed to hate material on Facebook and YouTube, with 21% becoming victims of such material. The author himself claims that cyber-bullying and cyber hate are on the rise but does not substantiate this with concrete evidence. Another survey suggested that 56% of young people in the UK said they had seen others bullied online. A survey on cyberstalking carried out in the UK with 349 respondents (of which 68% were women) found that the percentage of women and men saying they experienced harassment on social media was similar: 63.1% and 61.1% respectively" (Eckert, 2017).[10]

**Looking beyond studies focusing on the UK, there is more evidence on the extent of online trolling from other countries.** A number of studies could be identified estimating prevalence of online trolling and related behaviour in other European countries. Many more studies were identified estimating prevalence of cyberbullying in other countries than the UK or estimating prevalence without reference to specific geographies. Other studies report incidence rates of 31.4% for cyberbullying and 24.6 to 30.2% for cyber victimisation.

[8] The number of 2,752 is based on Twitter's claims that these are run by the Russian company Internet Research Agency (IRA). A further 1,062 such accounts were released by Twitter later on.

[9] https://democrats-intelligence.house.gov/uploadedfiles/exhibit_b.pdf

[10] This survey (known as the ECHO pilot study) was an online survey carried out by an anti-stalking NGO over six months between September 2010 and March 2011. The survey was hosted on the website of the Network Surviving Stalking, with an invitation provided to everyone who visited the website to take part. This method can be expected to skew respondents strongly towards those who have experienced stalking – and recognise it as such, and those who are more comfortable online. However, responses can be expected to be relatively representative of those who identify as stalking victims in the UK.

### Box: 2.5: Varying estimates of prevalence of online trolling in other European countries[11]

- Eckert (2017) reports on a study carried out by the European Union Agency for Fundamental Rights which in 2014[12] reported that "Of 48,000 women across the European Union, 11% said they experienced sexual online abuse on social media, in email or via SMS". The highest reported prevalence was found in Denmark and Sweden (18%), the lowest in Romania (5%). The UK prevalence was 13%. These numbers should be interpreted with caution and could be a reflection of underreporting and/or lower online penetration rather than actual prevalence of cyber-harassment. These figures are far lower than those of actual 'real-world' sexual harassment, where the reported EU average was 21% (25% in the UK).
- Another European study (Laftman et. Al, 2013) reported a 5% prevalence of being involved in cyber-bullying of which 4% were perpetrators and 2% were victims and perpetrators.
- Garaigordobil (2011) found that 20-50% of people experienced cyber-bullying victimisation, but only 2-7% were victims of severe cyber-bullying.
- Alonso et. al (2017) found that 10.3% of their sample had been implicated in cyber-bullying either as bullies or as victims.
- Jones et. al (2013) found the rate of online harassment nearly doubled between 2000 (6%) and 2010 (11%) in a sample of 4,561 individuals.
- Research by Seigfried-Spellar (2017) finds that 33% of youth experienced online harassment in 2016; 9% specifically on social networking platforms, and that 28% of adults engage in malicious activities towards strangers online, with 12% using controversial statements to start arguments.

A study carried out by Jenaro et. al (2017) reviewed several papers and reports estimates of international prevalence rates ranging from 9% to 72%[13], suggesting that it is very difficult to come up with a precise estimate of the prevalence of cyberbullying. A further study collected online data from 418 US residents of whom 5.6% reported enjoying engaging in trolling behaviour online. (Buckels et al, 2014).[14]

**Finally, another approach to estimating the prevalence of online trolling or cyberbullying is to extend the scope to examine aggressive online behaviour more generally (Rost et al, 2016).** The research in question found that of 532,197 comments on 1,612 online petitions, covering a range of subjects, published on the open petition Germany website[15] (which seems sufficiently large a sample to draw conclusions), 20.62% showed evidence of at least one aggressive expression (e.g. a swear

[11] For more information on each of these individual studies, please see annex 1

[12] http://fra.europa.eu/en/publication/2014/violence-against-women-eu-wide-survey-main-results-report

[13] The latter estimate is based on an anonymous online survey from 2008 of 12-17-year-olds in the US. Juvonen, Jaana and Elisheva F. Gross. 2008. Extending the School Grounds? — Bullying Experiences in Cyberspace. Available at: http://onlinelibrary.wiley.com/doi/10.1111/j.1746-1561.2008.00335.x/abstract;jsessionid=071C8C49D1C089024187959056CB2578.f04t04

[14] For this study, participants were recruited from Amazon's Mechanical Turk website (an online jobs platform matching companies/developers with individuals to manage surges in work requirements). Participants were paid $0.50 to complete a questionnaire, which included questions regarding trolling and online behaviours as part of a broader range of personality questions. Respondents were limited to those based in the USA. Amongst respondents, 42.3% were "non-commenters (i.e. did not engage in online discussions). When asked their "favoured activity when commenting", 23.8% of respondents like debating issues, 21.3% liked chatting generally, 2.1% said they liked making friends and 5.6% reported a preference for trolling. One of the major drawbacks of this survey was that recipients were asked to give their "favoured" activity – meaning that any crossovers, or multiple reasons for commenting, would be missed.

[15] This accounts for all comments on online petitions published on www.openpetition.de between its launch in May 2010 and July 2013.

word), with 9% showing two or more. Looking at the petitions themselves, aggression was also common, ranging from one aggressive expression (34%) with 28% showing more two or more; 47% of all petitions triggered controversial debates.

**The majority of papers in this section were published in scientific journals where they underwent a peer review.** However, a few pieces of grey literature (e.g. a European Parliament report, the Ofcom Report) are not peer reviewed. The largest samples used in studies referred to in this section are Ofcom's survey of 3,743 adults, the Pew Research Center survey of 3,217 individuals, and the Fundamental Rights Agency's survey of 48,000 women across the 28 EU Member States. The Rost et. al (2016) review of 532,197 comments on 1,612 online petitions in the UK is also worth mentioning.

**Box: 2.6: Conclusions - What is the prevalence of trolling and does this vary by type of social media platform?**

- Differences in methodology and definitions make it hard to draw parallels between studies;
- Furthermore, the quality of the evidence appears to be severely limited – either by surveys of unrepresentative groups, surveys of groups which are not clearly specified (explained), or surveys where questions regarding trolling are insufficiently clear.
- Estimates of the prevalence of online trolling vary widely (from 1% to 72%) of the sample population examined;
- Particular groups within society appear to be more affected by trolling. These tend to be groups that are more disempowered within society more generally, such as women, religious minorities, LGBTQ individuals etc;
- There is limited literature on the prevalence of trolling specifically, with more of a focus on cyber-bullying and abuse;
- No attempts were found to compare prevalence across social media platforms.

## 2.3 What is the profile of 'typical' trolls

**Establishing the profile of a 'typical' troll is difficult, in large part due to the numerous different definitions of trolls which can make it difficult to cross-compare studies.** There are also significant differences in the populations sampled (for example, studies looking at motivations tend to focus on interviews with relatively small groups on specific platforms, whereas larger surveys often include just one or two questions on trolling in a range of questions related to online behaviours). Nevertheless, we found a number of common traits being referenced in many of the papers, largely at the individual level but also some considering broader factors such as socio-economic indicators. Many of these traits are based on qualitative interviews and surveys, which rank character traits demonstrated by trolls against psychological indicators. This type of research should be treated very carefully, as it can be subjective and runs the risk of reflecting the researcher's own prejudices.

**A number of indicative psychological traits were consistently referred to in studies - most noteworthy is the 'dark tetrad' of narcissism, psychopathy, Machiavellianism, and everyday sadism** (see Buckels, 2014; Trapnell and Paulhus, 2014; Paulhus, 2014; Craker and March, 2016; Maltby, 2016; and March, 2017). Of these values, everyday sadism was found to be the most closely associated with online trolling in a literature review of social media and violence in Craker and March (2016), which surveyed self-described trolls, and Peterson and Denley (2017), a literature review of existing work on online violence.

### Box: 2.7: An explanation of the "Dark Tetrad"

The dark tetrad refers to four specific personality traits, which are outlined below:

- **Narcissism** - an excessive sense of self-love and self-admiration;
- **Psychopathy** - absence of empathy, lacking the emotional aspects of a conscience;
- **Machiavellianism** - a detached, calculating attitude regarding manipulativeness;
- **Everyday sadism** - (first coined in Buckels, Jones, and Paulhus 2013) refers to an enjoyment of cruelty in everyday culture, ranging from violent films and video games to incidents of police brutality.

A final common theoretical element is the 'Gyges effect' or 'Online Disinhibition Effect' – the idea that anonymity seemingly lowers self-awareness and inhibition, whilst encouraging a dissociation from the harassment, which has received robust empirical support (Hardaker 2015; Owens 2016).

**Whilst many of these papers are based on victims or bystander's conceptions of trolling, work directly with trolls themselves also suggests that these psychological traits are prevalent.** Zezulka (2016) used survey data from self-described cyberbullies and trolls to study the link between the two groups and found both types of cyber harassers were marked by low self-esteem, low conscientiousness and low internal moral values. Sest (2017), which also used a survey to attempt to study the demographics of trolls, suggests that the typical troll is male, with similar psychological traits to those of the "dark tetrad" but low affective empathy, allowing them to understand the emotional distress of others without empathising.

A study in March (2017) focused on users of the dating app Tinder found that high levels of impulsivity could predict trolling. However, this was only seen in cases of medium to high psychopathy[16], a personality disposition marked by a lack of conscience and by predation on others through manipulation and charm.

**A number of possible motivations for trolling are suggested across the papers: a need for attention, everyday sadism, low self-confidence, lack of empathy, and a desire for amusement.** In Jenaro *et al* (a study based on interviews with adults with learning disabilities), participants suggested that trolls were motivated to attack because their victims were different from them. On the other hand, Maltby (2017) refers to earlier studies which show 'griefers' (trolls in online games) are primarily motivated by amusement or entertainment tied to sadism, or gaining status in a trolling community, which suggests a plurality of opinions. There were also platform specific responses: Bratu (2017), for example, studied trolling on Facebook, and suggested that trolls aim for targets which are more likely to receive media coverage, and are motivated primarily to increase others' irritation. Craker and March (2016) also looked at Facebook posts. The paper (based on surveys with trolls) found that psychological traits and underlying motivation were intimately linked: a desire for self-serving goals such as entertainment had a positive association with the negative psychological traits considered above. Brown (2017) suggests that another motivation is community building online. Focusing on hate speech, it argues that engaging in online harassment is a way to cement or build further in-group status and create a kind of online community. Nevertheless, as an paper which reviews existing case studies, it is important to note that it is speculative rather than empirically supported.

**An alternative view on trolling suggested came from Cheng *et al* (2017), who proposed that rather than viewing trolls as atypical antisocial individuals, most of them are typical internet users.** In a study of the behaviour in the comment section of a major news site, trolling was linked to prior 'troll posts' by others, suggesting that it can provoke copycat behaviour, triggering similar moods and

[16] Psychopathy as used here should be interpreted with caution, as it is based on a 1980s model of psychopathic traits which appears to differ from modern clinical definitions.

behaviours to those exhibited by the troll in those exposed to them. Whilst the experiment was focused on a very specific form of online interaction (quite different to posts on a micro-blogging site or a dedicated social media platform), it nevertheless provides an alternative to the common concept of all trolls being aberrant.

**Another approach to defining motives for trolling comes from Cook, Shaafsma, and Antheunis (2017), which focuses on online gaming and used semi-structured interviews with 'griefers' to understand the triggers for trolling.** The most common of these was being trolled by another player, followed by boredom, a negative emotional state, and a desire to win (if not necessarily by the rules). The motivations tied to these triggers were personal enjoyment, revenge against those who had previously trolled them, and the thrill-seeking of partaking in an activity outside the games' norms or rules.

**It is important to note that specific kinds of trolling have different motivations, representing heterogeneity within the trolling community.** RIP trolling, the subject of Seigfried-Spellar and Chowdhury (2017), involves attacking memorial pages for the deceased. In the paper (which contained interviews with both RIP and non-RIP trolls), it became clear that individual trolls had different ideas regarding what constituted an appropriate target. Some RIP trolls only attacked memorial pages by so-called grief tourists, which the trolls believed did not know the deceased and were only posting for attention. This follows earlier work (particularly Phillips 2011), in which there was a difference of opinion between non-RIP trolls who thought attacking family members was inappropriate, and RIP trolls who claimed to feel no remorse and felt they were appropriate targets.

**Political trolling represents a rather different kind of online activity to both 'griefing' and RIP trolling, but evidence on motivation is weak.** Williams and Pearson (2016), which presented the findings of a panel of experts in Wales, suggested that anti-immigrant trolling there could be the result of trigger events such as terror attacks. Yet sincere political affiliation is sometimes lacking: Duggan (2014) provides some anecdotal evidence from victims that trolls belittled their opinion because they lacked a meaningful political stance of their own. This was discussed in relation to both right-wing and left-wing trolls, suggesting that trolls are able to adopt a position depending on the politics of the group in question. At the same time, given that this is anecdotal evidence, further research is needed.

A more organised form of political trolling involves using 'weaponised bots' (i.e. automated accounts, designed to send out identical or practically identical messages *en masse*), as discussed in Llewllyn (2018), which used Twitter scraping to analyse troll behaviour in the Brexit debate. Similar work carried out by Schafer, Evert, and Heinrich (2017) in relation to the Japanese election also found evidence of this behaviour, whilst pointing to divergence within the two different conservative campaigns seemingly involved in controlling the bots.

### Box: 2.8: Conclusions - What is the profile of 'typical' trolls

- Common psychological traits shared by trolls are described as the "dark tetrad" - narcissism, psychopathy, Machiavellianism and everyday sadism;
- Motivations identified include a need for attention and a desire for amusement;
- There is some evidence of copycat behaviour – those who are exposed to trolling behaviours are more likely to engage in it;
- There is some emerging evidence of different motivations and codes of conduct according to different sites/types of trolling. This evidence is limited, however, by a lack of broader studies comparing trolling across platforms.
- Motivations may differ according to the type of trolling (e.g. 'griefing' vs 'flaming' vs political trolling behaviours) and the cultural norms of the individual troll.

## 2.4 Is trolling a stepping stone/gateway to other negative behaviours?

**There is a limited literature dealing with trolling as a stepping stone to further negative behaviours, suggesting that it is difficult to assess the direction of association.** Of the papers examined, Zeulka and Seigfried-Spellar (2016) is the principal paper to deal with this issue in detail. In a comparison between the two prevalent kinds of online harassment – cyber-bullying, defined as involving victims known to the perpetrator, and trolling, defined as targeting strangers – similar psychological traits were detected. Whilst the paper did not conclude that one form of cyber harassment caused the other, the psychological overlap between the two suggests they are almost certainly linked in some cases.

**On a related note, several studies found links between online and offline bullying**. Myers and Cowie (2017) drew on secondary literature to show that cyberbullies tend to target individuals who are already bullied offline in traditional, face-to-face harassment. This is supported by Alonso (2017), which found a high co-occurrence between cyber-bullying and its non-digital counterpart, although it also fails to draw a conclusion on whether one triggers the other or if they develop in parallel. Evidence from a report for the European Parliament (2016) studying cyber-bullying amongst adolescents also strengthens this case: in some of the countries studied such as Germany, the link between being victimised online and becoming a perpetrator was high, with around a third of those who had cyberbullied having previously been cyberbullied. This fits with the idea in Myers and Cowie (2017) that cyber-bullying is a way of demonstrating power and dominance, and that cyberbullies are likely to have previously been victimised offline.

**The lack of research on whether trolling is simply a step to other negative behaviours (online or offline) is unfortunate.** That most of the texts available are either reviews of secondary literature or based on surveys of self-described trolls is also problematic – the former fails to generate additional data to analyse, and the latter has the risk of trolls responding mischievously to researchers.[17]

**Box: 2.9: Conclusions - Is trolling a stepping stone/gateway to other negative behaviours?**

- The literature available on this research question is very limited and there have been no systematic reviews assessing the quality of the secondary evidence;
- There is some evidence of overlap between cyber-bullying and online trolling;
- There is evidence of links between online and offline abuse;
- There is limited evidence that those who have previously been victimised online or offline also tend to become perpetrators of online abuse.

## 2.5 Can any differences be identified in the online and offline behaviour of trolls?

**Insufficient work has so far been carried out to link together trolls' presence online and offline.** The majority of papers assessed which deal with this subject are qualitative, based on textual analysis or interviews, or deal with secondary literature. While this is useful in providing some detailed insights into how trolling works in individual cases, findings would be further bolstered by strong quantitative research into online and offline behaviour. At present, studies tend to be based on anecdotal evidence that appears in the media or extrapolation from the apparent power of anonymity in studies such as Brown (2017) or Owens (2016). Secondly, and as a whole for studies of trolling, the criticism in Jane (2015) remains appropriate: there is a risk in research with trolls of

[17] Research into online gaming has identified a tendency to respond "mischievously", i.e. by providing extreme and untruthful responses that can dramatically effect estimates of phenomena such as variability in gender identity

scholars not critically questioning the motivations of respondents and the possibility for so-called "mischievous" responses.

**As discussed in point 2.3 above, Miller 2013 presents the case that online and offline identity are closely related and cannot be totally decoupled.** At the same time, this continuity of behaviour has to be considered alongside the importance of anonymity to trolls. As a result of the 'protection' of the screen, studies such as Owens (2016) show that users' inhibitions are lowered online due to the lower risks of discovery and consequences. This is explored best in Brown (2017), an observational qualitative research piece comparing online and offline hate speech. It provides five key affordances of the internet for online abuse. In addition to anonymity, there is the invisibility of perpetrators from targets and the attendant sense of distance, and the instantaneity of messages compared to in face-to-face or other offline bullying. Whilst the online and offline identities of the trolls have not changed, their harassment can be augmented through technology.

**Other papers suggest an overlap between online and offline behaviour in line with Miller's work.** Jenaro (2017) also argues that just as with offline harassment, victims are often chosen simply because they are different to the bullies. Jane (2015), a review of literature on online hostility, found that increasing levels of hostility online appears to correlate with threats of offline violence, such as bomb threats and poison sent by mail. It is also a noteworthy paper because whilst it does not generate new data or run any experiments, it recognises potentially serious flaws in research on troll behaviour: its three primary findings are that scholars have tended to underplay online hostility, to frame objections to hostility as prudish or sensitive, and to overlook the gendered nature of online harassment.

**Box: 2.10: Conclusions - Can any differences be identified in the online and offline behaviour of trolls?**

- The evidence suggests a continuity of identity between online and offline activity – e.g. hostility online appears to be linked to offline threats (such as hate mail, bomb threats etc)
- Nonetheless, anonymity is perceived as a "protective screen" which may lower inhibitions related to online activities

## 2.6 What is the profile of typical online trolling victims

The studies reviewed for this section were largely published in scientific journals, meaning they underwent some form of peer review. The exceptions concern a study published by the European Parliament and a report on an annual survey carried out in the UK. Most papers clearly lay out their argument, but the robustness of the evidence this rests on varies, reflecting different methodologies adopted ranging from quantitative analysis of online surveys over in-depth interviews informing case studies to literature reviews and legal analyses. The scatter-gun approach to trolling deployed so far, with researchers focusing on specific elements of trolling, also makes it difficult to build a clear picture of the impact and prevalence of trolling, as most findings rest on a limited evidence base – sometimes just a single study.

**The literature suggests that certain groups within society are particularly affected by online trolling and the related phenomenon of cyber-bullying.** These groups include women, ethnic minorities, religious groups, people with disabilities, and the LGBTQ community. Survey findings suggest that particular age (younger) and gender (women) groups are most closely associated with the experience of online harassment. The Pew Research Centre in the USA carries out a semi-regular survey which breaks down victims of online harassment into demographic groups, based on a probability-based, nationally representative panel in the USA. The 2014 survey was self-administered via the internet by 2,849 web users. The survey found that those aged 18-29 are more likely than any other demographic group to experience online harassment. Some 65% of young internet users have been the target of at least one of the six elements of harassment that were queried in the survey. Among those aged 18-24, the proportion is 70%. Young women in the 18-24 age group experience certain severe types of harassment at disproportionately high levels: 26% of these young women have been stalked online and 25% have been the target of online sexual harassment. The paper does not mention whether the same female survey respondents that stated they had experienced online stalking also experienced sexual harassment or physical threats. Hence, it is not possible to infer from the data that there is a correlation between these behaviours.

**In terms of gender, several papers were identified suggesting that women are targeted by online trolls and cyber-bullies more often than men.** Thus, one study (Peterson and Densley, 2017) found that girls made up 69% of victims in 2010. The study's findings are based on a literature review which seems to have considered a wide range of research. Another study (Eckert, 2017) suggests that likely targets of online abuse are women who identify as feminist on social media and other public forums. A further paper (Mantilla, 2013) investigated gender trolling, which also typically targets female victims, and specifically those who speak out about sexism in any form. In their study, Henry and Powell (2016) focused on online sexual abuse. They found that women and girls are the most likely, and most common, targets of online sexualised violence. Girls are more likely than boys to 'sext' as a result of coercion from male peers, and both women and girls are primary targets of revenge porn and cyberstalking. One study focusing on gender trolling in the UK (Girlguiding, 2016) found that women and girls experience being trolled for expressing their views and that this is often linked to threats of sexual violence. The study also found that cyber-bullying affects girls identifying as LGBTQ particularly.

**Braithwaite (2016) examined cross-national differences between social network behaviour among adolescents and young adults as well as whether various behavioural and demographic factors are associated with online harassment and victimisation.** The research covered the United States, Germany, Finland, and the UK. Within this study's sample, a few notable victim characteristics were identified. In the Finnish sample, females were 9.8% more likely to be a victim of online harassment. Also, previous research suggests (and this study confirms) that "those individuals having a large network of weak social ties have been shown to be more likely to experience harassment victimization online."

**Apart from gender and age, another way of developing a profile of typical trolling and cyber-bullying victims is to identify characteristic psychological traits.** One study (Alonso and Romero, 2013) found that cyber victims show higher levels of neuroticism, including high anxiety, hostility, depression and vulnerability to stress, high agreeableness (altruism and modesty), and higher levels of openness. These traits are based on the five-factor model of personality differences, pioneered by Goldberg (1990). As this is only one study, however, more research would be needed before any clear conclusions can be drawn.

**A few studies identified also highlighted marginal groups in society that were more likely to be trolled or cyberbullied.** These include Muslims, the LGBTQI (lesbian, gay, trans, queer and intersex) community, and people with disabilities. Barlow and Awan (2016) indicates that typical victims are usually members of groups that have faced socioeconomic inequality in the past, evidenced through reports from women and Muslims. It provides some wider observations about islamophobia and violence against women and girls online, before combining the two phenomena to demonstrate the particular vulnerability of female Muslims who, the authors argue, are more likely to be targeted due to their visual identifiable characteristics, such as a hijab or headscarf. This study is comprised of two auto-ethnographies, which the authors however acknowledge do not produce results which can be generalised. A qualitative interview study (Kilvington and Price 2017) found that men, particularly members of English football clubs, are subjected to racial trolling, which suggests that while gender is an important factor in identifying likely victims, perhaps common targets for online abuse are also figures in the public eye.

**Research by Jenaro et. Al (2017) looking at rates of cyber-bullying amongst adults aged 18-40 (269 participants) suggests that individuals with intellectual disabilities may find themselves at a higher risk of cyber-bullying.**[18] This is due to factors such as overusing Internet and mobile phones, carrying out unhealthy behaviours or being socially rejected for being different. This is based on a sample (269) from institutions for people with intellectual disabilities. In order to explore whether this initial finding is more applicable across the general population, a more broad-based survey would be needed, comparing rates of victimisation between adults with learning disabilities and those without learning disabilities.

**Conversely, a specific case comes from the US where a study found that that high-status/high-profile university students in the United States (e.g. athletes, student government officers) are often targeted by cyberbullies.** In addition, students who are involved in sororities and fraternities (known as "Greek life" in the United States) are disproportionately represented among cyberbullies and victims. Those who belonged to "Greek life" organisations were more frequent victims of humiliation and malice than non-members and perpetrated acts of public humiliation more often as well. The extent to which this occurs in other countries again remains unexplored research territory. Relationship difficulties, such as the break-up of a friendship or romance, were also linked to cyber-bullying at university; sexual orientation is also a significant factor that increases the risk of victimization (Myers and Cowie, 2017).

**Box: 2.11: Conclusions - What is the profile of typical online trolling victims**

- Gender, ethnic origin, religious identity, sexual identity, disabilities, visibility and possibly age all have the potential to impact the likelihood of becoming a victim of trolling.

## 2.7 What impact does trolling have on victims online and offline behaviour?

**There is very little quantitative evidence available to date regarding the impact of trolling on victims' online and offline behaviour.** Instead, the literature tends to focus on case studies, anecdotal evidence and self-reporting by victims of how they perceive trolling to have affected them. This can provide useful information in the form of detailed accounts regarding the impact of trolls on individuals' behaviour, but high-level study drawing on a broader and more robust body of data would be welcome to better understand if these impacts can be generalised across a population.

**A complication in seeking to assess impacts is that in much of the literature, no distinction is made between trolling and cyber-bullying – terms which are sometimes used interchangeably or used separately to refer to similar abusive online behaviours.** Online trolling appears to engender similar reactions in its victims to offline harassment (March, 2017; Park et. Al 2014), although it has been suggested that online behaviour may have a longer lasting and more pervasive impact than offline behaviour (Park et. Al 2014)**.** One suggested reason for this, as put forward in the Girls Attitude Survey of girls and women in the UK (Girlguiding 2016) is that while trolls tend to be anonymous, their actions are public and leave their victims exposed to large audiences.

[18] The sampling and data collection approaches taken to this study appear to be relatively robust, with efforts made to ensure participants were fully briefed and data was collected in a uniform manner. The data was also disaggregated by age and gender, to better ensure representativeness. However, as the reasons for cyberbullying were self-reported, there is a risk that these may have been misunderstood or misinterpreted by the participants in the study.

**Commonly reported symptoms amongst trolling victims include increased emotional distress and embarrassment, and increased risk of clinical or subclinical symptoms (such as those for depression, anxiety and posttraumatic stress disorder).** A number of the studies considered in this REA report similar behavioural impacts, including substance abuse, shame, humiliation, low self-esteem, paranoia, withdrawal from social life and detrimental impacts on personal relationships (for example with other family members, intimate partners) (Myers and Cowie, 2017; Englander, 2017; Maltby, 2017; Jenaro et. Al (2017); Duggan, 2015; Seigfried-Spellar, 2017; Pearson, 2016). In some more extreme circumstances, online trolling has also been linked to a higher risk of self-harm and suicide (Girlguiding 216; Maltby 2017; Pearson et. Al 2016). The evidence from these studies consists mainly of literature reviews, interviews, and qualitative and quantitative surveys. As there are overlapping themes in these findings, there is certainly some evidence that these problems are common among victims. There is some quantitative data to support these findings in the 2017 Ofcom report on Adults' Media use and attitudes, although further in-depth study would be welcome. Ofcom (2017) found that 44% of the 1,000+ survey respondents using social media agreed that they are put off from posting content because of the potential for abusive comments or responses. As a result, there has been a slight decline (compared to the previous year's survey) in trust in social media content. A national survey dedicated specifically to the prevalence and impact of trolling would be a useful tool in order to further develop these initial findings.

**In reviewing the literature as a whole, a number of factors emerge which appear to influence the impact of trolling upon its victims.** These include: the victim's perception of themselves; their perception of the troll; the interaction of these two perceptions in terms of the power dynamic they create between troll and victim; and the context in which the trolling takes place. Perceptions of power appear to be very important in terms of the magnitude of impacts on victims' behaviour. Victims who perceive trolling activities to be motivated by attention-seeking behaviours, for example, show increased resilience to online trolling, meaning that they do not experience the same negative well-being outcomes (Maltby 2016).

Victims who feel more directly threatened of for whom online trolling spills over into threats offline (for example, receiving rape threats, death threats, having personal details published online or threats to friends and family members), or who belong to groups which are more likely to be socially disempowered (such as Muslims or women, for example), tend to be more severely impacted (Maltby, 2016; Girlguiding, 2016; Barlow and Awan, 2016). An extreme example of this type of direct threat is revenge porn, when videos of sex acts are posted online to humiliate former sexual partners. This type of online abuse is particularly personal and is often associated with direct offline threats such as stalking and domestic violence. Impacts of this type of behaviour can be particularly severe: shame and humiliation; altered relationships with others; reputational damage; loss of employment prospects; victim blaming; withdrawal from social life and low self-esteem and paranoia" (Henry and Powell, 2016).

**Even when victims show relatively high levels of psychological resilience, they may still change their online behaviours in response to trolling.** In one study of female bloggers in Germany, Switzerland, the United Kingdom, and the United States, common responses to trolls included vetting comments in paper threads, self-censorship (keeping a "low profile online by not posting on "hot-button issues", toning down language or not promoting their blog to a broader readership) and, in some instances, removing themselves from social media altogether (Eckert, 2017). Further evidence comes from an opinion survey of 293 self-identified female gamers recruited through online forums, blogs and social media sited (Fox and Tang, 2017)[19], which examined women who had experienced sexual harassment in online video games. Those who responded to this survey reported

[19] Responses only refer to completed surveys, and data was disaggregated by age. It is unclear what types of questions were used, so it is hard to gauge to what extent these might have been leading.

withdrawing from the spaces in which harassment occurred, and a level of indifference from the gaming organisations involved to the abuse they received.

**Furthermore, even those who are resilient to trolling themselves may change their behaviours in response to threats directed at others (for example, family members or those within their communities).** One well-known example is Anita Sarkeesian, who received a significant amount of online abuse after publishing a series of online videos in the US entitled "Tropes versus women in video games". Ms Sarkeesian chose to speak out against the abuse she received personally, rather than practicing self-censorship. She nonetheless cancelled a talk at Utah State University due to the institution receiving a threatening e-mail from an anonymous source that the individual would conduct a "Montreal Massacre-style attack" against attendees and students and staff at the nearby Women's Centre (Braithwaite, 2016; Barlow and Awan, 2016).

**Another important finding was the effect trolling had on a victim's home life.** In Barlow and Awan (2016), Awan remarks how the constant barrage of hateful, Islamophobic comments left her emotionally drained, and her family begged her to cease writing opinion pieces. Indeed, withdrawal can be encouraged by those within a victim's support network. Mantilla's (2013) evidence review of GamerGate and other instances of gender-trolling features several examples of this. Kathy Sierra moved homes, for example, after trolls sent packages to her home to demonstrate that they knew where she lived.

**Box: 2.12: Conclusions - What impact does trolling have on victims online and offline behaviour?**

- Psychological impacts identified include increased emotional distress and embarrassment, and increased risk of clinical or subclinical symptoms (such as those for depression, anxiety and post-traumatic stress disorder);
- Behavioural impacts identified include substance abuse, withdrawal from social life, detrimental impacts on personal relationships, and - in extreme cases - self harm and suicide;
- The evidence suggests that groups that are more socially disempowered may be more likely to be more affected by trolling;
- There is some evidence that the impact of trolling on victims' behaviour is linked to their perception of the troll, with stronger resilience if the troll is perceived as a nuisance (lacking in power over the victim) rather than a direct threat.

## 2.8 Can any practical methods be identified to challenge trolling? How effective have past interventions been?

Robust evaluation evidence based on large and representative surveys and datasets, well-targeted questions, and appropriate and transparent sampling methods on specific interventions to challenge trolling is limited, with a tendency to focus on small-scale experiments to flag trolls. While the researchers' aim is to embed these within popular social media platforms, none of these have been integrated so far. This is to say, social media platforms still rely on their own automatic flagging mechanisms to report and challenge trolls. As such, no robust studies to show the impact of these interventions have been identified.

**The literature reviewed highlights two types of approaches to combat online trolling.** The first relies on automatic mechanisms that rely on algorithms or other forms of artificial intelligence, whereas the second involves human-based interventions or 'soft' measures that require deliberate action by individuals in response to specific occurrences, such as online-community group decisions to ignore trolls. Papers advocating human-based interventions identify community-led strategies on a number of social media platforms or blogs. Those that argue in favour of automatic mechanisms either suggest their own metrics based on quantitative evaluative studies or explain how existing online platforms deal with trolls. A number of papers also look at the law as a means to challenge trolls and/or make recommendations on how to challenge trolling in a number of contexts.

**Authors that experimented with automated methods tend to focus on detecting and codifying trolling behaviours online that feed algorithms.** Some of these automated tools are generally scales that rate the severity of trolling or the vulnerability to trolling. The end goal is to embed these scales (e.g., recurrence within the same individual, platform) into social media platforms so bots or trolls can be detected automatically. For example, Wikipedia's method to detect vandalism or the modification of posts made in bad faith relies on a machine learning algorithm, which depends on a selection of features that are inputs to the algorithm (Adler et. Al, 2011). Another study looking at Wikipedia's strategy developed a theory to distinguish the behaviour of typical users, as opposed to benign users (Kumer et. Al, 2015). The study tried to detect vandals, or benign users, before an automatic trolling detection system flagged them to Wikipedia by identifying vandalism patterns. These methods were based on algorithms that essentially enabled distinguishing between vandals and non-vandals (or typical users). This indicates that automated detection of trolls involves analysing and understanding normal patterns of user behaviour, in contrast to focussing on the 'pathological' behaviour of troll accounts.

**A specific research tool was employed in one case to challenge trolling bots, entitled BotOrNot (Grimme et. Al, 2017).** This tool tries to establish the probability that a Twitter account is automated by learning patterns on the account's data, network, behavioural timing, online friendships and content. The results of this experiment show that the system does not work to detect hybrid social bots, a particular type of bot that can get itself followers, advocate certain ideas and generate automatic messages in order to bypass detection systems. The next big challenge is to generate detection mechanisms that can expose real human behaviour, over bots that can model social behaviour. The Troll Filtering Process, which has been tested to exploit the cognitive and affective information to assess a level of troll-ness in each post, then classify users and prevent trolls from emotionally hurting people within social networks (Cambria et. Al, 2010). The aim of this exercise is to replace 'soft' methods by embedding this scale on social media websites to reassure users.

**An alternative method proposed within the literature is Troll Vulnerability Rank - a metric based on the amount of trolling activity that followed after a post was published (Tsantarliotis et. al, 2016).** The goal is to then introduce a troll prediction, where posts can acquire TVRank value if their post is likely to receive a lot of troll attention. This pro-active method differs from automated methods that merely detect and remove trolls after they post inflammatory content, in that it anticipates troll activity before it takes place.

**One method that was identified as effective in countering troll posts is 'disemvowelling', or the practice of removing vowels from harassing comments to make it unintelligible to the reader (Shaw, 2013).** This practice was identified on a blogging platform called Jezebel and is deemed to be an effective way to cut down the amount of trolling. Jezebel is a blog geared towards women, research findings suggest that without 'disemvowelling' trolls would probably be more successful in derailing feminist conversations. Online platforms often also rely on banning certain users based on their disruptive behaviours.

**Many scholars highlight online-community strategies, or 'soft' methods to ignore or discredit trolls.**

Some identified the method of 'icing out' trolls by deliberately excluding users (Bratu, 2017). Some online communities warn users with messages such as DNFTT (Do Not Feed The Trolls) but it has been found that this method is not effective in dealing with trolls (Cambria et. Al, 2010). This is further confirmed by a study suggesting that DNFTT is not as effective as banning users and using 'disemvowelling' (Shaw, 2013). Automatic responses, especially from social networks, can also be inefficient if they do not recognise a particular type of trolling, e.g. gender-trolling on the platform Yik Yak (Eckert, 2017). Another social/soft method identified within feminist online communities, which involved drawing attention to the troll by making visible the discourse trolls are using to derail the conversation. This strategy is entitled 'speaking of the unspeakable' through 'heaping' and accumulation through, for instance, the use of hashtags #mencallmethings on Twitter (Shaw, 2013). On Twitter a similar initiative has been recorded, where users deflect hate speech by posting more positive messages and encouraging hashtags (Shaw, 2013).

**The literature also makes other recommendations to challenge trolling such as developing a website for victims of online hate crime to discuss their experiences (Shaw, 2013).** Clinicians asking their patients on their internet use to guide them towards preventative sites (Westerlund et. Al, 2015) is also suggested along with more intervention methods for therapists (Englander, 2017), psychological treatment aimed at cyber addiction in general (Jenaro et. al, 2017), and an EU-wide definition of cyber-bullying (European Parliament, 2016). The papers also make suggestions on how cyber-bullying can be tackled on university campuses, but these solutions have not been tested. In Sweden, there are online hate insurances that assist victims in handling the legal aspects of asserting their rights when subject to cyber-bullying in general (Enarsson and Naarttijärvi, 2016).

**In addition, many papers look at what the law says and can do to protect victims, as well as what cases have taken place in a given country to combat trolling.** In the UK and some other countries, the law does little to protect victims of trolling (Enarsson and Naarttijärvi, 2016). Research by Myers and Cowie (2017) explores the legal situation in the UK and identifies what legal options are available to victims of trolling. The study highlights the perspectives of both lawyers and practitioner psychologists on trolling and cyberbullying, finding that although the UK has legal requirements for schools to have a behaviour policy that includes measures to prevent all forms of bullying, but there is no centralised law or legal requirement for universities to have anti-bullying policies in place. Furthermore, cyberbullying is not recognised as a crime in the UK, and there is no single law specifically for such behaviours. The exception to this is revenge-porn, which has been a criminal offence since 2015. Lawyers interviewed for this study cautioned against passing one overarching Act on trolling and cyberbullying. The sample size consisted of only 9 interviewees, however, implying that a larger consultation on legal remedies to trolling would be welcome.

**The introduction of a trolling magnitude scale, an objective measure of the severity of online trolling in a given situation, has been suggested to enable trolls to be prosecuted in an objective way by authorities (Owens, 2016).** However, it is not clear how the scale would work in practice, what the criteria would be to evaluate the severity of trolling and how effective it would be in practice.

**Box: 2.13: Conclusions - Can any practical methods be identified to challenge trolling? How effective have past interventions been?**

- Two main responses to online trolling are identified: "soft" community-led strategies and "hard" automated measures based on algorithms to identify trolls';
- Much of the literature focuses on specific technical solutions developed by individual researchers. However, the evidence regarding the effectiveness of these solutions in real-world situations is very limited.
- Other methods involve increasing victim's psychological resilience and awareness, for example through psychological interventions aimed at trolls and victims;
- There is little evidence of legal protection offered for victims of trolling in the literature reviewed. In February 2018, the UK government asked the Law Commission to review the laws around offensive communications and assess whether they provide the right protection to

victims online[20];

- There is no evidence of nationwide government-level interventions to challenge trolling. Initiatives have mainly been undertaken by non-governmental organisations, particularly within the field of academia;
- Overall, there is no robust evaluative research on specific interventions and their impacts.

[20] For more information on this, see the Prime Minister's speech on the matter in February 2018 or the press release published by the Law Commission

# 3. Conclusions and Recommendations

## 3.1 Overall Conclusions

**The term trolling is used in the literature to cover a broad range of abusive online activities, with no uniform definition being applied.** The majority of papers examined include some definition of trolling, or at least define a sub-category of trolling, e.g. gender-trolling, flaming, etc. Trolling is a relatively recent phenomenon and one which appears to be evolving, from a slightly anarchic form of activity carried out by individuals towards a more automated form of online abuse, as has been seen for example in the targeting of various high profile figures and activities (such as elections) by automated “bots”.

With regard to developing a working definition of trolling, some common parameters emerge from the rapid evidence assessment:

**Box: 3.1: Defining Trolling**

- **Context:** trolling is an activity which is carried out online and is associated with activities where debate is encouraged (e.g. social media platforms, online forums, discussion and comment threads, online gaming chat groups etc).
- **Anonymity**: there is usually a perception of anonymity associated with the perpetrator, or the relationship is distant, for example in the case of attack on disempowered social groups. Counter to this, the victim is vulnerable to exposure as trolling is a public act.
- **Activity:** trolling involves posting off-topic material, inflammatory or confusing messages.
- **Motivation:** trolling can be used to create disruption and discord, to provoke a response from individuals or groups of users, or as a silencing tool to discourage other internet users from getting involved with additional online discussion. Trolling may be undertaken for amusement or in order to cause harm to specified targets.

Furthermore, when the troll is perceived to be powerful, the actions appear to have a far more significant impact in terms of the harm on the victim. When the troll is perceived as powerless, their actions tend to be viewed as a nuisance rather than posing a direct threat to the victim.

**The Rapid Evidence Assessment on trolling underlines a lack of robust, inter-disciplinary study on trolling, in terms of both impact and prevalence.** This is important both to capture the complexity of trolling itself and to overcome the rather fragmented approach to studying trolling identified in this evidence assessment. While the term is often used in academic literature, it lacks clear definition and appears to be used more as a “keyword” to draw attention to study on a plethora of subjects related to online activity than as a developed subject of academic study in its own right.

## 3.2 Key Findings on Research Questions

### 3.2.1 What is the prevalence of trolling and does this vary by type of social media platform?

**Estimates of prevalence identified within the literature vary widely, according to the definition of trolling used, the population studied (including factors such as age, gender, physical location and social group).** Prevalence is also measured according to different metrics (number of perpetrators, number of victims and level of activity), which can make it hard to draw meaningful comparisons between studies. The literature demonstrates a clearer focus on cyber-bullying (possible because this is a better defined term and therefore easier to measure), with limited attempts to quantify the prevalence of online trolling specifically. Furthermore, the quality of the evidence is not strong.

**A principal methodological weakness shared by many studies providing estimates of the prevalence of online trolling is that they rely on surveys which could be affected by selection biases.** None of the studies were found to provide confidence intervals for their estimates of prevalence, which would allow some understanding of how small or large the actual prevalence might be in the population.

**With regard to the UK, one of the most useful estimates of prevalence of online trolling comes from the 2017 Ofcom report on Adults' Media use and attitudes.** Two data sources have been used to inform the report: a survey of 1,846 adults aged 16 and over, and results from Ofcom's Technology Tracker based on another survey of 3,743 adults aged 16 and over. According to the study, 1% of Internet users in the UK had been trolled online at least once over the past 12 months. This goes up to 5% for respondents aged 16-24.

**Looking beyond studies focusing on the UK, there is more evidence on the extent of online trolling from other countries.** A number of studies could be identified estimating prevalence of online trolling and related behaviour in other European countries. Many more studies were identified estimating prevalence of cyberbullying in other countries than the UK or estimating prevalence without reference to specific geographies. Other studies report incidence rates of 31.4% for cyberbullying and 24.6 to 30.2% for cyber victimisation.

### 3.2.2 What is the profile of 'typical' trolls

**Establishing the profile of a 'typical' troll is difficult, in large part due to the numerous different definitions of trolls.** Nevertheless, we found a sizeable literature pertaining to common traits shared by trolls, largely at the individual level but also some considering broader factors such as socio-economic indicators. A number of indicative psychological traits were consistently referred to in studies - most noteworthy is the 'dark tetrad' of narcissism, psychopathy, Machiavellianism, and everyday sadism.

**A number of possible motivations for trolling are suggested across the material that has been reviewed**: a need for attention, everyday sadism, low self-confidence, lack of empathy, and a desire for amusement. An alternative view on trolling suggested came from Cheng et al (2017), who proposed that rather than viewing trolls as atypical antisocial individuals, most of them are typical internet users. It is important to note that specific kinds of trolling have different motivations, representing heterogeneity within the trolling community.

### 3.2.3 Is trolling a stepping stone/gateway to other negative behaviours?

**There is a limited literature dealing with trolling as a stepping stone to further negative behaviours, suggesting that it is difficult to assess the direction of association.**

**The evidence on whether trolling is simply a step to other negative behaviours (online or offline) is limited.** That the majority of the texts available are either reviews of secondary literature or based on surveys of self-described trolls is also problematic – the former fails to generate additional data to analyse, and the latter has the risk of trolls responding mischievously to researchers.

### 3.2.4 Can any differences be identified in the online and offline behaviour of trolls?

**There is some suggestion in the literature reviewed that online and offline identity are closely related and cannot be totally decoupled.** At the same time, this continuity of behaviour has to be considered alongside the importance of anonymity to trolls. As a result of the 'protection' of the screen, studies show that users' inhibitions are lowered online due to the lower risks of discovery and consequences.

**However, insufficient work has been carried out to link trolls' presence online and offline.** Furthermore, the majority of papers in the section are qualitative, based on textual analysis or interviews, or deal with secondary literature. While this is useful in providing some detailed insights into how trolling works in individual cases, findings would be further bolstered by strong quantitative research into online and offline behaviour.

### 3.2.5 What is the profile of typical online trolling victims

**The literature suggests that certain groups within society are particularly affected by online trolling and the related phenomenon of cyber-bullying.** These groups include women, ethnic minorities, religious groups, people with disabilities, and the LGBT community. Survey findings suggest that particular age (younger) and gender (women) are most closely associated with the experience of online harassment. In terms of gender, several papers were identified suggesting that women are targeted by online trolls and cyber-bullying more often than men.

**Apart from gender and age, another way of developing a profile of typical trolling and cyber-bullying victims is to identify characteristic psychological traits.** One study examined found that cyber victims show higher levels of neuroticism, including high anxiety, hostility, depression and vulnerability to stress, high agreeableness (altruism and modesty), and higher levels of openness. A few studies identified also highlighted marginal groups in society that were more likely to be trolled or cyberbullied.

### 3.2.6 What impact does trolling have on victims online and offline behaviour?

**There is very little quantitative evidence available to date regarding the impact of trolling on victims' online and offline behaviour.** Instead, the literature tends to focus on case studies, anecdotal evidence and self-reporting by victims of how they perceive trolling to have affected them.

**There does appear to be some evidence of continuity between online and offline behaviour, both for trolls and victims of trolling.** For trolls, the literature suggests this lies in a correlation between their tendency to be abusive online and offline. Specifically, there appears to be evidence of a spill-over between online abuse in some context and offline abuse such as hate mail, death threats and stalking. The literature on psychological traits of trolls suggests that those underlying tendencies elicit problematic behaviour both online and offline, and that is why such a correlation can be found. Little direct evidence is available to support this hypothesis as yet, however. For victims, trolling may lead to a reduction in engagement online and social withdrawal offline. In situations where online and offline abuse are happening in combination, the impact on the victim tends to be more significant.

**A complication in seeking to assess impacts is that in much of the literature, no distinction is made between trolling and cyber-bullying – terms which are sometimes used interchangeably or used separately to refer to similar abusive online behaviours.** In the same way as with cyber bullying, online trolling appears to engender similar reactions in its victims to offline harassment.

**Commonly reported symptoms amongst trolling victims include increased emotional distress and embarrassment, and increased risk of clinical or subclinical symptoms (such as those for depression, anxiety and posttraumatic stress disorder).** A number of the studies identify similar behavioural impacts, including substance abuse, shame, humiliation, low self-esteem, paranoia, withdrawal from social life and detrimental impacts on personal relationships (for example with other family members, intimate partners).

**Even when victims show relatively high levels of psychological resilience, they may still change their online behaviours in response to trolling**. Furthermore, even those who are resilient to trolling themselves may change their behaviours in response to threats directed at others (for example, family members or those within their communities).

### 3.2.7 Can any practical methods be identified to challenge trolling? How effective have past interventions been?

**The literature reviewed highlights two types of approaches to combat online trolling.** The first relies on automatic mechanisms that rely on algorithms or other forms of artificial intelligence, whereas the second involves human-based interventions or ‘soft’ measures that require deliberate action by individuals in response to specific occurrences, such as online-community group decisions to ignore trolls.

**Many papers look at what the law says and can do to protect victims, as well as what cases have taken place in a given country to combat trolling.** There is a consensus among legal scholars that the victim is the one who needs to take responsibility to deal with harassments and threats online. This is particularly the case since in the UK and some other countries where the law does little to protect victims of trolling.

**Very few papers were found which assessed effectiveness or efficiency of reporting mechanisms on social media platforms, suggesting there is little-to-no research on how effective this is.** This is probably because companies are reluctant to release data. An exception to this was one study on Wikipedia which tested the effectiveness of a suggested algorithm against a manual sample in identifying “troll posts”. Another paper on social bots on Twitter tried to measure the percentage of success of reporting hybrid social bots.

**Furthermore, there is limited information available on how major online platforms address trolling and the effectiveness of the strategies currently being used.** A review of transparency reports published by Facebook, Twitter and Google provided little clear insight into trolling activities on this platform and instead reported requests filed by the UK government for criminal investigation, shutting down accounts etc. This lack of publicly available information may have affected the methods used by scholars to assess prevalence of trolling and effectiveness of strategies to counter it.

**Attempts to combat trolling identified in the research have tended to look at a combination of top-down approaches such as vetting of comments before publication or the imposition of codes, or bottom-up approaches where users work together to neutralise the activities of trolls.** There is also an increased interest in the use of automated IT tools to identify and neutralise trolls, although these have yet to be tested in practice. Little specific legal protection is currently offered for victims of trolling, and there is little evidence of coordinated efforts to tackle the issue at national level in the UK.

**Given the gaps in what is known, more research is needed to gain a clearer understanding of what trolling is, how it is affecting UK society and what needs to be done to effectively counter its negative impacts.** Nonetheless, the literature reviewed as part of this research has highlighted some common findings and areas of agreement amongst researchers working in these areas, as well as pointing to some interesting hypotheses, which could be explored in further study.

## 3.3 Recommendations

Based on our review, suggestions are made below for future research and policy actions on this topic:

- A UK-wide survey is needed to better understand the prevalence and impact of trolling, based on a clear definition of trolling.
- More research into cross-platform comparisons of trolling activities and attempts to counter them would provide valuable insights.
- The literature is characterised by fragmentation between different disciplines. An inter-

disciplinary approach covering legal, psychological, technological and other aspects of trolling would be welcome.

There is some evidence of a link between online trolling behaviours and traditional bullying. More research into this aspect should be considered as it would inform future prevention strategies (i.e. whether “soft” preventative approaches such as the provision of psychological support should be provided or whether “hard” technological solutions are better suited to dealing with the problem).

# Appendix A: Overview of Literature

**Name of paper:** *I refuse to respond to this obvious troll*

**Paper Reference:** Hardaker, C. (2015). ‘I refuse to respond to this obvious troll’: an overview of responses to (perceived) trolling. Corpora, vo. 10, No. 2 : pp. 201-229

**Brief summary of study characteristics:**

Computer mediated communication (CMC) can be a fertile ground for trolling, especially since it gives a sense of impunity and/or suppresses empathy. This paper seeks to address the question, ‘How do users respond to (perceived) trolling?’. The answer to this is elaborated through the creation of a working taxonomy of response types, drawn from 3,727 examples of user discussions and accusations of trolling which were extracted from an eighty-six million word Usenet corpus.

______________________________________________________________________________

**Name of paper:** *Bullying at University: The Social and Legal Contexts of Cyber-bullying Among University Students*

**Paper Reference:** Myers, C.A. and Cowie, H., 2017. Bullying at University: The Social and Legal Contexts of Cyber-bullying Among University Students. Journal of Cross-Cultural Psychology, 48(8), pp.1172-1182.

**Brief summary of study characteristics:**

The paper only tackles cyber-bullying in general, without going into too much detail of what type of cyber-bullying/ which country/ which type of university. No mention of trolling but rather covers different types of online bullying. In this paper, they consider the social and cultural contexts that either promote or discourage cyber-bullying among university students. Finally, the implications for policies, training, and awareness raising are discussed along with ideas for possible future research in this under researched area. The paper specifically looks at university students, and to some extent the situation in the UK (situation of cyber-bullying, but also policy recommendations for UK universities to tackle cyber-bullying).

______________________________________________________________________________

**Name of paper:** *Cyber-Violence: Towards a Predictive Model, Drawing upon Genetics, Psychology and Neuroscience*

**Paper Reference:** Owens, T. (2016). Cyber-Violence: Towards a Predictive Model, Drawing upon Genetics, Psychology and Neuroscience. International Journal of Criminology and Sociological Theory, 9 (1). pp. 1-11

**Brief summary of study characteristics:**

Paper contends that the definition of cyber violence should include a wider spectrum of hostile and aggressive behaviour in cyber space. The paper explores a predictive model of cyber violence drawn on the multifactorial analysis favoured in the Genetic-Social meta-theoretical framework.

______________________________________________________________________________

**Name of paper:** *Flaming? What flaming? The pitfalls and potentials of researching online hostility*

**Paper Reference:** Jane, E. A. (2015) Flaming? What flaming? The pitfalls and potentials of researching online hostility. Dordrecht: Springer Science and Business Media. 65-87

**Brief summary of study characteristics**:

This paper identifies several critical problems with the last 30 years of research into hostile communication on the internet and offers suggestions about how scholars might address these problems. This paper is a review of the literature, thus is focused on researchers and does not reveal trolling/flaming practices, nor does it provide practical solutions. Author also introduces their own term: e-bile, which is comparable to flaming trolling and refers to: “vitriolic discourse notable for its hostile affect, explicit language, and stark misogyny”. Main conclusion, the author advocates a scholarly shift which conceptualises online hostility as a broad field of inquiry whose horizon is an ethical one.

---

**Name of paper:** *Implicit theories of online trolling: Evidence that attention-seeking conceptions are associated with increased psychological resilience*

**Paper Reference:** Maltby J. (2017). Implicit theories of online trolling: Evidence that attention-seeking conceptions are associated with increased psychological resilience. British Journal of Psychology: 448-466.

**Brief summary of study characteristics**:

Three studies were made for this research to investigate people’s conceptions of online trolls, particularly conceptions associated with psychological resilience to trolling. Study (1): participants rated characteristics of online trolling – 5 group factors were highlighted; (2) participants evaluated hypothetical profiles of online trolling messages to establish the validity of the five factors; (3) introduced a 20- item ‘Conceptions of Online Trolls scale’ to examine the extent to which the five group factors were associated with resilience to trolling. Results indicated that viewing online trolls as seeking conflict or attention was associated with a decrease in individuals' negative affect around previous trolling incidents.

---

**Name of paper:** *Is it all part of the game? Victim differentiation and the normative protection of victims of online antagonism under the European Convention on Human Rights*

**Paper Reference:** Enarsson, T. and Naarttijärvi, M., 2016. Is it all part of the game? Victim differentiation and the normative protection of victims of online antagonism under the European Convention on Human Rights. International review of victimology, 22(2), pp.123-138.

**Brief summary of study characteristics**:

This paper analyses the issue of online antagonism (defined as antagonistic harassment, defamation, insults and threats online) and the positive obligations of states to counter such antagonism under the European Convention on Human Rights, from a legal and victimological perspective. The paper is based on the Swedish legislative context. The paper aims to clarify to what extent victims can expect states to provide legal remedies when faced with online antagonism. This is done through an analysis of the protection of freedom of expression under the ECHR in relation to the scope of positive obligations under paper 8.

Most of the paper specifically looks at the situation of Sweden, which is not entirely relevant to subject matter/scope.

**Name of paper:** *Regulating Cyber-Racism*

**Paper Reference:** Mason, G. (2017) Regulating Cyber-Racism. Melbourne: Melbourne University Law Review Association Inc. 284-340

**Brief summary of study characteristics:**

The paper examines the current legal and regulatory terrain around cyber-racism in Australia. The analysis exposes a gap in the capacity of current regulatory mechanisms to provide a prompt, efficient and enforceable system for responding to harmful online content of a racial nature. Drawing on recent legislative developments in tackling harmful content online, it considers the potential benefits and limitations of key elements of a civil penalties scheme to fill the gap in the present regulatory environment. It argues for a multifaceted approach, which encompasses enforcement mechanisms to target both perpetrators and intermediaries once in- platform avenues are exhausted.

***Name of paper:*** *Social Bots: Human-Like by Means of Human Control?*

**Paper Reference:** Grimme C et al. (2017). Social Bots: Human-Like by Means of Human Control? Big Data 5(4): 279-293.

**Brief summary of study characteristics**:

This paper aims to define what **social bots** are and provides an overview of how social bots actually work (taking the example of Twitter) and what their current technical limitations are. The paper then discusses how bot capabilities can be extended and controlled by integrating humans into the process and reason that this is currently the most promising way to go in order to realize effective interactions with other humans.

***Name of paper:*** *Still 'Searching for Safety Online': collective strategies and discursive resistance to trolling and harassment in a feminist network*

**Paper Reference:** Frances Shaw (2013). Still 'Searching for Safety Online': collective strategies and discursive resistance to trolling and harassment in a feminist network. Fibreculture Journal (22): 92-107.

**Brief summary of study characteristics**:

This paper examines the discursive responses that participants in a network of feminist blogs developed to handle trolling in their community. Internet communities develop strategies to deal with trolls in their networks. In particular, participants provide instructions and guidance to support each other to deal with trolls and harassment and engage in intra-community discussion about the significance or insignificance of trolls. The author explores the different practices that feminist use to resist keeping silent from these trolls. This research looked only at Australian bloggers. The research was done through 20 bloggers from 2009-2010 (potentially outdated).

***Name of paper:*** *The Inexorable Shift Towards an Increasingly Hostile Cyberspace Environment: The Adverse Social Impact of Online Trolling Behavior*

**Paper Reference:** Bratu, S. (2017) The Inexorable Shift Towards an Increasingly Hostile Cyberspace Environment: The Adverse Social Impact of Online Trolling Behavior. Woodside: Addleton Academic Publishers. 88-94

**Brief summary of study characteristics**:

The purpose of this paper is to gain a deeper understanding of the moral consequences of online trolling behaviour, the peculiarities of the troll space, the effect of socio-cultural and technological settings on online trolling, and the personality features and incentives of persons that get involved in Facebook trolling. This paper is an observation piece, in which the author draws from other research to make their own points on the motivations behind trolling and what are common traits of trolls.

---

***Name of paper:*** *Troll vulnerability in online social networks*

**Paper Reference:** Tsantarliotis, P et. Al (2016). Troll vulnerability in online social networks, IEEE/ACM ASONAM 2016, August 18-21, 2016, San Francisco, CA, USA.

**Brief summary of study characteristics**:

This paper takes a novel approach to the trolling problem: goal is to identify the targets of the trolls, so as to prevent trolling before it happens. The aim is to predict whether a post is vulnerable to trolling. As such, they define a novel troll vulnerability metric of how likely a post is to be attacked by trolls, and then construct models for predicting troll-vulnerable posts, using features from the content and the history of the post. Experimental data comes from Reddit.

---

***Name of paper:*** *What is so special about online (as compared to offline) hate speech?*

**Paper Reference:** Brown, A., 2017. What is so special about online (as compared to offline) hate speech?, Ethnicities, p.1468796817709846.

**Brief summary of study characteristics**:

There is a growing body of literature on whether or not online hate speech, or cyber- hate, might be special compared to offline hate speech. This paper aims to both critique and augment that literature by emphasising a distinctive feature of the Internet and of cyberhate that, unlike other features, such as ease of access, size of audience, and anonymity, is often overlooked: namely, instantaneousness. The paper is an observational qualitative research piece on online vs. offline online hate.

---

***Name of paper:*** *"You Need to Be Sorted Out With a Knife": The Attempted Online Silencing of Women and People of Muslim Faith Within Academia*

**Paper Reference:** Barlow, C., Awan, I., 2016. "You Need to Be Sorted Out With a Knife": The Attempted Online Silencing of Women and People of Muslim Faith Within Academia. Social Media + Society 2, 2056305116678896.

**Brief summary of study characteristics**:

This study hones in on gendertrolling against female academics who speak out against misogyny and Islamophobia on social media platforms. By taking such a position, they are criticized not based on what they said, but on their identities, as trolls attempt to silence them with threats of rape, stalking, and real-world violence, on top of insults regarding their bodies and faces. This paper examines two auto-ethnographic reports and how they exemplify the impact such trolling has on its victims.

---

***Name of paper:*** *"Defining Cyber-bullying"*

**Paper Reference:** Englander, E. (2017). Defining Cyber-bullying. Pediatrics 140(2): 148-151

**Brief summary of study characteristics**:

The paper defines cyber-bullying and determines whether there is a difference between it and traditional bullying. The implicit argument here is that cyber-bullying is a distinct form of bullying and may inflict more psychological and emotional harm than traditional bullying. The main keyword here, "cyber-bullying" is defined in the first paragraph using different definitions from two studies. The evidence presented is taken from previous literature and is quite convincing since it appears grounded in past studies and observations. This paper only provides definitions and solutions, and does not discuss trolling. The closest it comes to distinguishing trolling from cyber-bullying is "online harassment". It also could do with some more literature on the impact such behaviour has on victims.

---

***Name of paper:*** *"Fighting for recognition: Online abuse of women bloggers in Germany, Switzerland, the United Kingdom, and the United States"*

**Paper Reference:** Eckert, S., 2017. Fighting for recognition: Online abuse of women bloggers in Germany, Switzerland, the United Kingdom, and the United States. New Media and Society, p.1461444816688457.

**Brief summary of study characteristics**:

The paper is about feminist bloggers in specific nations, and the online abuse they receive for voicing their opinions. One argument the author poses is that understanding the gravity of online abuse requires us to consider offline incidents that are products of or reactions to a user's online posts, and that at the moment these incidents are not taken as seriously as they should be. These offline incidents should not be considered as separate from the online sphere, as they can implicate others related to the victims of online abuse, and are situated within a context of digital inequalities, gender inequality and other social constructs. The author uses the Pew Research Center's definition of "online abuse", and quickly defines blogging to justify their interview sample. However, the author does not define troll until midway through the study.

The evidence in this study is mostly experiential, as the author believes in-depth interviews with feminist bloggers would provide richer data and firsthand accounts of their experiences with online abuse. Other relevant evidence is cited from other studies. Although the sample size (109) is not terribly small, especially for a qualitative interview study, perhaps having even numbers of interviews in each country would ground the findings in more equal comparisons. This study was based off of 19 interviews in Switzerland and the U.K., compared to 34 in Germany and 37 in the United States.

---

***Name of paper:*** *"Gendertrolling: Misogyny Adapts to New Media"*

**Paper Reference:** Mantilla, K., 2013. Gendertrolling: Misogyny adapts to new media. Feminist Studies, 39(2), pp.563-570.

**Brief summary of study characteristics**:

The essay's primary argument is that gendertrolling is distinct from more generic forms of trolling, and therefore deserves special attention. Another underlying argument is that gendertrolling is perhaps worse than other forms of trolling because of the extent to which trolls attempt to silence and undermine their victims. The key terms—"trolling", "lulz", and "gendertrolling"—are defined in the introduction.

---

***Name of paper:*** *"Histories of Hating"*

**Paper Reference:** Shepherd, T., Harvey, A., Jordan, T., Srauy, S. and Miltner, K., 2015. Histories of hating. Social Media+ Society, 1(2), p.2056305115603997.

**Brief summary of study characteristics**:

The paper presents a transcript of a roundtable discussion. The principal focus is on #GamerGate, but broader topics of online abuse are discussed. There are several implicit arguments, as there is a plurality of opinions among the multiple authors. For example, one author suggests that there is a larger underlying conflict on the Internet: a shift in normative values. This shift is characterised by free speech movements, a nostalgia for the Internet's libertarian, Wild West days when anyone could say anything without content moderation because everyone was anonymous. Another author suggests that trolls' behaviour is misconstrued in the online social world. Their playful "just a meme" or "just trolling" excuses serve as thin veils for their hate. This plays into the paper's broader argument that trolling is performative. Counter arguments regarding trolling as a whole and how to regulate it were fully considered. For example, some argue for the psychological basis of trolling, others assert that trolling is a cultural issue.

---

***Name of paper:*** *"It's About Ethics in Games Journalism? Gamergaters and Geek Masculinity"*

**Paper Reference:** Braithwaite, A., 2016. It's about ethics in games journalism? Gamergaters and geek masculinity. Social Media+ Society, 2(4), p.2056305116672484.

**Brief summary of study characteristics**:

The main argument of the paper is that #GamerGate is beyond a toxic discussion about gender and ethics in the gaming industry; it is "an articulation of technology, privilege, and power" in which GamerGaters are using the scandal to assert their masculinity and reclaim territory they believe is under siege. Another argument is that this phenomenon exemplifies how social media can expose misogyny on a massive scale. In terms of estimating prevalence, the belief that up to 400,000 users participated in #GamerGate is based on views for relevant YouTube videos, which the author acknowledges is generous since a user can view a video multiple times, or one view can equal multiple people crowded around a laptop screen.

---

***Name of paper:*** *"Sexual Violence in the Digital Age: The Scope and Limits of Criminal Law"*

**Paper Reference:** Henry, N. and Powell, A., 2016. Sexual violence in the digital age: The scope and limits of criminal law. Social and Legal Studies, 25(4), pp.397-418.

**Brief summary of study characteristics**:

The study looks at different forms of technology-facilitated sexual violence and its regulation in Australia. There are several arguments put forth. The primary argument is that Australian law is ill-equipped to deal with the gendered harm resulting from technology-facilitated sexual violence (TFSV) due to its slow pace, outdated policies, and lack of effective enforcement throughout the country. The key words are well defined and help the reader better understand the argument. Another important argument is that TFSV is a gendered phenomenon. The counter-argument is that men and boys are also subjected to TFSV and the effects are likewise significant. The authors acknowledge this in the introduction and in the section dedicated to cyberstalking. Although men are also stalked, their cases are less common compared to women's, and they are more likely to be perpetrated by a stranger or acquaintance rather than a former partner.

---

***Name of paper:*** *"Social Tie Strength and Online Victimization: An Analysis of Young People Aged 15-30 Years in Four Nations"*

**Paper Reference:** Keipi, T., Kaakinen, M., Oksanen, A. and Räsänen, P., 2017. Social tie strength and online victimization: An analysis of young people aged 15–30 years in four nations. Social Media+ Society, 3(1), p.2056305117690013.

**Brief summary of study characteristics**:

This study is about cross-national differences between social network behaviour among adolescents and young adults in the United States, Germany, Finland, and the UK; as well as whether various behavioural and demographic factors are associated with online harassment and victimisation. . The main argument here is that having a large network of online friends and strong identification with online communities are positively related to victimization and harassment, while the younger one is, the more likely they are to have experienced negative online behaviours

---

***Name of paper:*** *"Tackling Social Media Abuse? Critically Assessing English Football's Response to Online Racism"*

**Paper Reference:** Kilvington, D. and Price, J., 2017. Tackling Social Media Abuse? Critically Assessing English Football's Response to Online Racism. Communication and Sport, p.2167479517745300.

**Brief summary of study characteristics**:

The study looks at online racism within the world of English football. There are several arguments in this paper, even though it is an examination of how sports institutions are addressing racist and discriminatory social media posts. One is that "the problem of sports-related racism on social media is a worldwide one with social media organisations transcending national legal boundaries." (2) Another is that any response to online racism from a national sports institution has an international stage due to international fanbases. A third, more general implicit argument is that the Internet has not lived up to its utopian vision of an egalitarian space where identities are hidden behind screens and anonymous avatars. The main argument here is that even if offline racism at football matches has decreased, racism towards players, clubs and fans has increased online, especially on social media.

---

***Name of paper:*** *"Women's experiences with general and sexual harassment in online video games: Rumination, organizational responsiveness, withdrawal, and coping strategies"*

**Paper Reference:** Fox, J. and Tang, W.Y., 2017. Women's experiences with general and sexual harassment in online video games: Rumination, organizational responsiveness, withdrawal, and coping strategies. New Media and Society, 19(8), pp.1290-1307

**Brief summary of study characteristics**:

This study looks at harassment in video games, and how women cope with such behaviours.

The implicit argument is that women are especially prone to harassment on online gaming platforms. The arguments presented are all logical, and echo similar arguments from gaming studies. The introduction includes signposting of the three variables being investigated, then moves on to a review of online videogaming culture, definition of sexual harassment, and where these two worlds collide. The authors then discuss their experimental method and results before drawing conclusions. The counter argument is that men may be more likely to experience some forms of online harassment, "such as being insulted or embarrassed by other players" (1291).

---

***Name of paper:*** *Adults' Media use and attitudes*

**Paper Reference:** Adults' media use and attitudes, 2017, Ofcom

**Brief summary of study characteristics**:

The study is mostly interesting in that it provides an indication of prevalence based on a survey covering adults' media use and attitudes. The survey is carried out on an annual basis, and this study presents results of its 12th iteration.

---

***Name of paper:*** *Aggressors and Victims in Bullying and Cyber-bullying: A Study of Personality Profiles using the Five-Factor Model*

**Paper Reference:** Alonso C et al. (2017). Aggressors and Victims in Bullying and Cyber-bullying: A Study of Personality Profiles using the Five-Factor Model. The Spanish Journal of Psychology.

**Brief summary of study characteristics**:

This study aims to examine the link between bullying and cyber-bullying, and thus sheds light on the difference between online and offline behaviour of trolls. It uses a five-factor model (a well-established model for personality traits) to identify characteristics associated with different roles implicated in cyber-bullying, namely perpetrators and victims. Information was collected through a survey of 910 adolescents aged 12-19 years.

---

***Name of paper:*** *Cyber violence: What do we know and where do we go from here?*

**Paper Reference:** Peterson, J and Densley, J. (2017) Cyber violence: What do we know and where do we go from here?. Aggression and Violent Behavior, Volume 34, Pages 193-200

**Brief summary of study characteristics**:

A literature review looking at the link between social media and violence, including prevalence rates, typologies, and the overlap between cyber and in-person violence (on- and offline behaviour). Also explores the causal or contingent role of social media in violent offending.

---

***Name of paper:*** *Cyber-bullying among adults with intellectual disabilities: Some preliminary data.*

**Paper Reference:** Jenaro C et al. (2017). Cyber-bullying among adults with intellectual disabilities: Some preliminary data. Research in Developmental Disabilities.

**Brief summary of study characteristics**:

This study is based on young people with mental disabilities and assesses to what extent they risk experiencing cyber-bullying. The findings are based on a sample of 269 participants from Chile, Mexico, and Spain.

---

***Name of paper:*** *Cyber-bullying among young people*

**Paper Reference:** European Parliament. (2016). Cyber-bullying among young people.

**Brief summary of study characteristics**:

Building on a combination of desk research and legal analysis, this report provides an overview of the extent, scope and forms of cyber-bullying in the EU and looks at the profile of victims and perpetrators. The study illustrates the legal and policy measures on cyber-bullying adopted at EU and international levels and across EU member states. Outlines variety of definitions of cyber-bullying and presents successful practice on how to prevent and combat it. Focuses on psycho trolling. The study focuses on people younger than 18. Included because of 16-18 age bracket, but to be treated with caution.

---

***Name of paper:*** *For Whom the Bell Trolls: Troll Behaviour in the Twitter Brexit Debate*

**Paper Reference:** Llewelyn, C. et. al (2018). For Whom the Bell Trolls: Troll Behaviour in the Twitter Brexit Debate.

**Brief summary of study characteristics**:

Investigated the activities of 2,752 accounts Twitter believed to be controlled by Russian operatives, which were used to influence the US Presidential election. In the absence of a similar list of operatives active within the debate on the 2016 UK referendum on membership of the European Union (Brexit) the behaviour of these American Election focused accounts was studied in the production of content related to the UK-EU referendum.  There is a focus on political trolling.

---

***Name of paper:*** *Girls' Attitude Survey*

**Paper Reference:** Girls' attitudes survey, 2016, GirlGuiding

**Brief summary of study characteristics**:

Each year, Girlguiding's Girls' Attitudes Survey takes a snapshot of what girls and young women think on a wide range of issues. This major survey, now in its eighth year, canvasses the opinions of over 1,600 girls and young women aged 7 to 21, inside and outside guiding across the UK.

---

***Name of paper:*** *Hate Crime and Bullying in the Age of Social Media*

**Paper Reference:** Williams, M., Pearson, O., 2016, Hate Crime and Bullying in the Age of Social Media, Cardiff University

**Brief summary of study characteristics**:

This is a conference paper reporting on discussions of a panel of 100+ experts from across industry, public and third sectors. The All Wales Hate Crime Project highlighted the emerging problem of cyberhate and cyber bulling via social media through interviews with victims. This includes negative practices such as the production and contagion of misinformation and antagonistic and prejudiced commentary.

___

***Name of paper:*** *Online Harrassment*

**Paper Reference:** Duggan, M., 2014, Online Harrassment, Pew Research Centre

**Brief summary of study characteristics**:

Reports on findings from a probability-based, nationally representative panel (in the US). This survey was conducted May 30 – June 30, 2014 and self-administered via the internet by 2,849 web users, with a margin of error of plus or minus 2.4 percentage points.

___

***Name of paper:*** *Trolling on Tinder (and other dating apps): Examining the role of the Dark Tetrad and impulsivity*

**Paper Reference:** March, E. (2017) Trolling on Tinder (and other dating apps): Examining the role of the Dark Tetrad and impulsivity. Elsevier Ltd.

**Brief summary of study characteristics**:

Study seeks to explore trolling on Location-Based Time Dating apps, correlating participants' sex to known trolling traits. Sought to predict preparation of trolling on these apps. Found no sex differences, but that impulsivity can predict trolling perpetration if the individual has medium-to-high psychopathy levels. The study findings are based on a sample of 357 Australians sourced from the community of a dating app. Focuses on playtime trolling.

___

***Name of paper:*** *Online Harassment 2017*

**Paper Reference:** Duggan, Maeve, et al. "Online harassment 2017." The Pew Research Center.(11 July 2017). Retrieved September 8 (2017): 2017.

**Brief summary of study characteristics**:

Study seeks to demonstrate how widespread online harassment is in America. Found that four in ten Americans have suffered harassment based on a sample of 4,248 US adults. 14% of those surveyed had been targeted because of political views, 9% because of physical appearance, and 8% because of race or gender.

___

***Name of paper:*** *Death and Lulz: Understanding the personality characteristics of RIP trolls*

**Paper Reference:** Seigfried-Spellar, K.C. (2017) Death and Lulz: Understanding the personality characteristics of RIP trolls. Purdue University: First Monday

**Brief summary of study characteristics**:

Study of RIP trolls (i.e. trolls who attack memorial pages to the deceased on platforms like Facbook), using interviews with both RIP trolls and other trolls to show differences within the trolling community.

---

***Name of paper:*** *Constructing the cyber-troll: Psychopathy, sadism, and empathy*

**Paper Reference:** Sest, N. (2017) Constructing the cyber-troll: Psychopathy, sadism, and empathy. Australia: Elsevier Ltd.

**Brief summary of study characteristics**:

Use of a quantitative questionnaire to study demographics of trolling, as well as the psychological traits which underlie trolling online.

---

***Name of paper:*** *Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions*

**Paper Reference:** Cheng, J., e.t. al, 2017, Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions, US National Library of Medicine, National Institute of Health, 1217-1230

**Brief summary of study characteristics**:

Analysis of online news discussion board, which points to mood and the context of the discussion as having greater predictive power than previous history of trolling (arguing against previous history suggesting that certain personality traits/motivations are central to trolls acting as they do).

---

***Name of paper:*** *Differentiating Cyberbullies and Internet Trolls by Personality Characteristics and Self-Esteem*

**Paper Reference:** Zezulka, L. (2016) Differentiating Cyberbullies and Internet Trolls by Personality Characteristics and Self-Esteem. Farmville: Association of Digital Forensics, Security and Law. 7-25

**Brief summary of study characteristics**:

Paper studying individual differences and self-esteem between self-reported cyberbullies and/or internet trolls, to attempt to assess prevalence of the behaviours in relation to each other.

---

***Name of paper:*** *Digital Social Norm Enforcement: Online Firestorms in Social Media.*

**Paper Reference:** Rost K et al. (2016). Digital Social Norm Enforcement: Online Firestorms in Social Media. PLoS ONE 11(6).

**Brief summary of study characteristics**:

Application of social norm theory to online aggression, supported by analysis of a German online petition site (http://www.openpetition.de) and the comments there. Results point towards a greater degree of aggression showed by non-anonymous users compared to anonymous users, particularly when there is intrinsic motivation (i.e. a desire to enforce societal norms).

---

***Name of paper:*** *Implicit theories of online trolling: Evidence that attention-seeking conceptions are associated with increased psychological resilience*

**Paper Reference:** Maltby, J. (2016) Implicit theories of Online trolling: Evidence that attention-seeking conceptions are associated with increased psychological resilience. Leicester: Wiley Subscription Services. 448-466

**Brief summary of study characteristics**:

Report based on three studies focusing on conceptions of online trolling, specifically psychological resilience to trolling. Concludes that use of an implicit theories approach can improve the measuring of conceptions of trolling, and points towards a resilience strategy against trolling.

---

***Name of paper:*** *Japan's 2014 General Election: Political Bots, Right-Wing Internet Activism, and Prime Minister Shinzo Abe's Hidden Nationalist Agenda*

**Paper Reference:** Schäfer F et al. (2017). Japan's 2014 General Election: Political Bots, Right-Wing Internet Activism, and Prime Minister Shinzō Abe's Hidden Nationalist Agenda. Big Data 5(4): 294-309.

**Brief summary of study characteristics**:

The study is focused on identifying and analysing the behaviour of social bots (automated account) on Twitter, specifically in relation to right-wing users with an affinity for Shinzo Abe before and after the 2014 Japanese general election.

---

***Name of paper:*** *Perceived Severity of Cyber-bullying: Differences and Similarities across Four Countries.*

**Paper Reference:** Palladino BE. (2017). Perceived Severity of Cyber-bullying: Differences and Similarities across Four Countries. Frontiers in Psychology.

**Brief summary of study characteristics**:

Study of Estonian, Italian, German, and Turkish adolescents (48.2% female, aged 12 to 20, with a mean age of 14.49), and considering the perceived severity of cyber-bullying across countries. Findings suggest a similar structure is apparent in all of them, with imbalance of power, anonymity, intentionality, and repetition being central to affecting the severity of cyber-bullying (with some national differences).

---

***Name of paper:*** *The dark side of Facebook®: The Dark Tetrad, negative social potency, and trolling behaviours*

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

**Brief summary of study characteristics**:

Study specifically of Facebook, and how the negative qualities of the Dark Tetrad (narcissism, Machiavellianism, psychopathy, and sadism), combined with social reward affect trolling behaviour. The study found that psychopathy and sadism explained trolling on Facebook to a point, but that the motivation of a negative social reward had the greatest impact on predicting whether an indivdiaul would engage in antisocial trolling behaviour.

---

***Name of paper:*** *Toward a Taxonomy of Dark Personalities*

**Paper Reference:** Paulhus, D.L., 2014. Toward a taxonomy of dark personalities. Current Directions in Psychological Science, 23(6), pp.421-426.

**Brief summary of study characteristics**:

A psychological review of the Dark Tetrad, an underlying set of personality traits which have been linked to trolling behaviour in a number of papers.

---

***Name of paper:*** *Trolling in online discussions: from provocation to community building*

**Paper Reference:** Hopkinson, C. (2013). Trolling in online discussions: from provocation to community building. Brno studies in English. 2013, vol. 39, iss. 1, pp. 5-25

**Brief summary of study characteristics**:

A paper discussing how users of British newspapers websites define trolling, which points towards the impact of discussion topics on framing. It also considers trolling behaviour on an attack on one of these websites, as well as considering the generative effects of trolling.

---

***Name of paper:*** *Under the bridge: An in-depth examination of online trolling in the gaming context*

**Paper Reference:** Cook, C., Schaafsma, J. and Antheunis, M., 2017. Under the bridge: An in-depth examination of online trolling in the gaming context. New Media and Society, p.1461444817748578.

**Brief summary of study characteristics**:

A study of trolls within video gaming communities, focused on their own perceptions of what counts as trolling, their motivations, and the communities response to trolling (as perceived by trolls), as discovered through semi-structured interviews.

---

***Name of paper:*** *Do Not Feel The Trolls*

**Paper Reference:** Cambria E, Chandra P, Sharma A and Hussain A (2010) Do not feel the Trolls

**Brief summary of study characteristics:**

Study provides a definition of trolling. Adopts a methodology combining sentic computing, opinion mining and sentiment analysis to detect trolls and hence prevent web-users from being emotionally hurt.

---

***Name of paper:*** *Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features*

**Paper Reference:** Adler, B.T., de Alfaro, L., Mola-Velasco, S.M., Rosso, P., and West, A.G. (2011). Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features. In CICLing '11: Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics, LNCS 6609, pp. 277-288

**Brief summary of study characteristics:**
Study presents results of an effort to integrate a spatio-temporal analysis of metadata (STiki), a reputationbased system (WikiTrust), and natural language processing features to detect Wikipedia vandalism. Argues that this new method is far more effective in identifying Wikipedia vandalism.

---

***Name of paper:*** *Beyond vandalism: Wikipedia trolls.*

**Paper Reference:** Shachaf, P., and Hara, N. (2010). Beyond vandalism: Wikipedia trolls. Journal of Information Science, 36, 357–370.

**Brief summary of study characteristics:**
This study looks at the behaviour and motivations of Wikipedia trolls. It has been included here, because it was referenced by studies identified in the initial review although it is outside of the REA scope.

---

***Name of paper:*** *VEWS: A Wikipedia Vandal Early Warning System, University of Maryland*

**Paper Reference:** Kumar, S., e.t. al, 2015, VEWS: A Wikipedia Vandal Early Warning System, University of Maryland

**Brief summary of study characteristics:**
This study presents a new model to classify Wikipedia vandal behaviour in Wikipedia users with an accuracy of 95-90%, using a sample of 35,000 Wikipedia users (including both a black list and a white list of editors) and 770,000 edits.

---

***Name of paper:*** *Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions*

**Paper Reference:** Hardaker, C. (2010). Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. Journal of Polymer Research, 6, 215–242.

**Brief summary of study characteristics:**
This study is technically outside of the scope of the REA, but an analysis has been included as it is referred to by multiple studies. Argues that computer-mediated communication provides anonymity that may encourage a sense of impunity and freedom from being held accountable for inappropriate online behaviour. Argues that a definition of trolling should be informed first and foremost by user discussions. Taking examples from a 172-million-word, asynchronous computer-mediated communication corpus, four interrelated conditions of aggression, deception, disruption, and success are discussed. Finally, a working definition of trolling is presented.

---

***Name of paper:*** *Trolls just want to have fun*

**Paper Reference:** Jonason, P. K., Lyons, M., Bethell, E., and Ross, R. 2014. Trolls just want to have fun

**Brief summary of study characteristics:**
The study presents findings from two online surveys (1,215 respondents/users) on personality traits and Internet commenting styles A key finding is that trolling correlates most strongly with sadism, followed by psychopathy and Machiavellianism.

***Name of paper:*** *The bad boys and girls of cyberspace: How gender and context impact perception of and reaction to trolling.*

**Paper Reference:** Fichman, P., and Sanfilippo, M. (2015). The bad boys and girls of cyberspace: How gender and context impact perception of and reaction to trolling. Social Science Computer Review, 33, 163–180.

**Brief summary of study characteristics:**
This study focuses on how gender impacts online trolling and to understand if men and women perceive and react differently to trolls and if trolls' gender impact perception of and reaction to trolls and their motives. Results indicate that men and women react differently to online trolling, and their perceptions of the impact of trolling on online communities vary; men and women identify different motivations for similar behaviours in different communities, and they both perceive that men and women trolls differ in their behaviour and motivation. The study is based on a survey of 100 participants.

***Name of paper:*** Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation

**Paper Reference** Howard, P and Bradshaw, P, (2017). Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation. Computational Propaganda Research Project, 2017.12, 1-37.

**Brief summary of study characteristics**:
Working paper on government-sponsored social media manipulation. Compares organisations across 28 countries (including democracies and authoritarian regimes), looking at kinds of messages, valences and communication strategies used. Many different countries employ significant numbers of people and resources to manage and manipulate public opinion online, sometimes targeting domestic audiences and sometimes targeting foreign publics. Looking across the 28 countries, every authoritarian regime has social media campaigns targeting their own populations, while only a few of them target foreign publics. In contrast, almost every democracy in this sample has organized social media campaigns that target foreign publics, while political-party-supported campaigns target domestic voters. Over time, the primary mode for organizing cyber troops has gone from involving military units that experiment with manipulating public opinion over social media networks to strategic communication firms that take contracts from governments for social media campaigns.

***Name of paper*** Cyber Violence

**Paper Reference:** Owens, T. (2017). Cyber Violence (Crime, Genes, Neuroscience and Cyberspace). London: Palgrave Macmillan, pp 103-114

**Brief summary of study characteristics:**

An attempt to provide a broader framework under which to categorise acts of online aggression. Cyber violence can be regarded as behaviour by an actor which takes place online and which is hostile and aggressive, and which may also be offensive, indecent, obscene or of a menacing character. The victims can be of any background with regard to age, gender, ethnicity, sexuality or social class. Such cyber violence can be found within both the 'known' parts of cyberspace, the social media sites, forums, chat rooms and 'normal' webpages indexed by conventional search engines, and the 'dark net', which 'has come to mean the encrypted world of Tor Hidden Services', where users cannot be traced, and cannot be identified (Bartlett 2015: 3). Also attempts to provide a framework for predicting cyber violence.

---

***Name of paper*** Bullying and Cyberbullying: Their Legal Status and Use in Psychological Assessment

**Paper Reference:** Samara, M., 2017, Bullying and Cyberbullying: Their Legal Status and Use in Psychological Assessment, Int J Environ Res Public Health, 14(12):1449

**Brief summary of study characteristics:**

This is a qualitative research that includes interviews with five practitioner psychologists and four lawyers in the United Kingdom (UK). Thematic analysis revealed three main themes. One theme is related to the definition, characteristics, and impact of bullying and cyberbullying and the need for more discussion among the psychological and legal professions. Another theme is related to current professional procedures and the inclusion of questions about bullying and cyberbullying in psychological risk assessments. The third theme emphasised the importance of intervention through education. Two key messages were highlighted by the lawyers: ample yet problematic legislation exists, and knowledge will ensure legal success. The study recommends the necessity of performing revisions in the clinical psychological practices and assessments, and the legal policies regarding bullying and cyberbullying. In addition to improving legal success, this will reduce bullying prevalence rates, psychological distress, and psychopathology that can be comorbid or emerge as a result of this behaviour.

---

***Name of paper*** Offender–victim relationship and offender motivation in the context of indirect cyber abuse

**Paper Reference** Vakhitova, Z., Webster, J., Alston-Knox, C., Reynald, D. and Townsley, M., 2018. Offender–victim relationship and offender motivation in the context of indirect cyber abuse: A mixed-method exploratory analysis. International Review of Victimology, p.0269758017743073.

**Brief summary of study characteristics:**

Study focusing on how indirect cyber abuse (circulating documents about a victim online) reflects motivation for attacker (primarily instrumental)

---

***Name of paper*** Olympic Trolls: Mainstream Memes and Digital Discord?

**Paper Reference** Tama Leaver. (2014). Olympic Trolls: Mainstream Memes and Digital Discord? Fibreculture Journal (22): 215-232.

**Brief summary of study characteristics:**

Explores definitions of online trolling and argues that wider forms of online abuse are being subsumed under that term, and that online trolling has been 'mainstreamed' in that it has become a widespread online communication strategy as part of everyday digital discourse.

***Name of paper*** Online Harrassment and Cyber Bullying

**Paper Reference** Dent, J and Strickland, P (2017). Online Harrassment and Cyber Bullying. London: UK Parliament

**Brief summary of study characteristics:**
Online harassment and cyber bullying can take a wide variety of forms including:"trolling" (sending menacing or upsetting messages); identity theft; "doxxing" (making available personal information); cyber stalking. It can affect adults and children. Some argue that online bullying amongst school children is more pervasive than face to face bullying, because it can follow a child home after school, and from one school to another. The problem of online abuse of Members of Parliament has also been highlighted in recent months, particularly of female and ethnic minority MPs.

***Name of paper*** Self-control in Online Discussions: Disinhibited Online Behavior as a Failure to Recognize Social Cues

**Paper Reference** Voggeser, B., e.t. al, 2017, Self-control in Online Discussions: Disinhibited Online Behavior as a Failure to Recognize Social Cues, Frontiers in Psychology, 8:2372

**Brief summary of study characteristics:**
An online experiment examining the role of self-control in recognizing social cues in the context of disinhibited online behaviour (e.g., flaming and trolling). Illustrates that self-control failure may manifest itself in a failure to recognize social cues. The finding underlines the importance of self-control in understanding disinhibited online behaviour: Many instances of disinhibited online behaviour may occur not because people are unable to control themselves, but because they do not realize that a situation calls for self-control in the first place.

## Papers rejected following the first sift

***Name of paper:*** *Bullying and Cyber-bullying in Adolescence*

**Paper Reference:** Genta M et al. (2009). Bullying and Cyber-bullying in Adolescence. Carocci.

**Brief summary of study characteristics**: Unable to obtain access to this book

***Name of paper:*** *"Case study of posts before and after a suicide on a Swedish internet forum"*

**Paper Reference:** Westerlund, M. Hadlackzy, G. and Wasserman, D. (2015). "Case study of posts before and after a suicide on a Swedish internet forum". The British Journal of Psychology. 207(6): 476-482.

**Brief summary of study characteristics**:

The study considers how social media users react to explicit suicide threats. Although not entirely clear, an implicit argument could be that suicidal individuals who post online do not receive the support or encouragement they need to stay alive, and that negative comments affect them more than positive or sympathetic ones. "Troll" and "Flashback" (the social media platform examined in this study) were defined to set the stage. Types of comments in the analysis section were also well-defined.

This study doesn't appear to pertain to our research. There may still be useful takeaways, but it is certainly not the most informative paper in terms of victim behaviour, since the case in question was not a suicide due to being trolled.

---

***Name of paper:*** *"Online Stigma Resistance in the Pro-Ana Community"*

**Paper Reference:** Yeshua-Katz, D., 2015. Online stigma resistance in the pro-ana community. Qualitative health research, 25(10), pp.1347-1358.

**Brief summary of study characteristics**:

This paper is completely irrelevant to our research. This paper is about online support, not trolling. The only way it would be relevant is if the stigmatized individuals in the pro-ana community were trolled and seeking refuge but that is not the case here.

---

***Name of paper:*** *Digital Teens and the 'Antisocial Network': Prevalence of Troublesome Online Youth Groups and Internet trolling in Great Britain*

**Paper Reference:** Bishop, J (2014). Digital Teens and the 'Antisocial Network': Prevalence of Troublesome Online Youth Groups and Internet trolling in Great Britain. International Journal of E-Politics, 5(3), 1-15

**Brief summary of study characteristics**:

This paper looks at data collected from subjects aged 50 in three UK regions (n=150 to 161 – ), which includes young people who are not in education, employment or training (NEETs). The data shows that these NEETs factors are usually not the cause of Internet trolling but that it is in fact the areas with high levels of productivity, higher education and higher intelligence but that report lower perceptions of quality of life that correlate with Internet trolling. The paper focuses on playtime trolling. The methodology used is not explained and the paper is highly speculative, so it has been excluded from this assessment.

---

***Name of paper:*** *The art of trolling law enforcement: a review and model for implementing 'flame trolling' legislation enacted in Great Britain (1981–2012)*

**Paper Reference:** Bishop, J., 2013, The art of trolling law enforcement: a review and model for implementing 'flame trolling' legislation enacted in Great Britain (1981–2012), International Review of Law, Computers and Technology, Vol. 27: 3

**Brief summary of study characteristics**:

This paper reviews the legislation enacted in the UK parliament between 1981 and 2012 to deal with 'flame trolling', or electronic message faults more generally. The paper presents a framework for identifying and addressing this phenomenon. The paper traces the political response to the phenomenon over the years and generally argues in favour of a cautious approach making sure prosecution does not undermine freedom of speech. The paper is poorly argued and poorly grounded, so it was left out of this study.

---

***Name of paper:*** *Future Identities: Changing identities in the UK – the next 10 years - What is the relationship between identities that people construct, express and consume online and those offline?*

**Paper Reference:** Danny, M., 2013, What is the relationship between identities that people construct, express and consume online and those offline?, University College London, published: Government Office for Science

**Brief summary of study characteristics**:

A report for the Government Office of Science, which combines theoretical understandings of identity from early digital studies research to the mid-2000s as a kind of extended literature review, before offering some predictions for the future of identity. While the research is interesting, it does not respond to the needs of this study

---

***Name of paper:*** *"Uh. . . . not to be nitpicky,,,,,but…the past tense of drag is dragged, not drug.": An overview of trolling strategies*

**Paper Reference:** Hardaker, C. (2013). "Uh. . . . not to be nitpicky,,,,,but…the past tense of drag is dragged, not drug.": An overview of trolling strategies. Journal of Language Aggression and Conflict, Volume 1, Issue 1, pages: 58 –86.

**Brief summary of study characteristics**:
Draws on 3,727 examples of user discussions and accusations of trolling from an eighty-six million word Usenet corpus in order to better understand trolling strategies and motications. Initial findings suggest that trolling is perceived to broadly fall across a cline with covert strategies and overt strategies at each pole. A working taxonomy of perceived strategies that occur at different points along this cline is developed and a working definition of trolling is provided. Left out of review as it seems to duplicate with another paper published by the same author and the provenance is not clear.

---

***Name of paper:*** *Stakeholder Perceptions of Cyberbullying Cases: Application of the Uniform Definition of Bullying.*

**Paper Reference :** Moreno MA et al. (2018). Stakeholder Perceptions of Cyberbullying Cases: Application of the Uniform Definition of Bullying. The Journal of Adolescent Health

**Brief summary of study characteristics**:
Applies the uniform definition of bullying to cases of cyberbullying. Found that uniform definition only applies in 56% of cases of cyberbullying. Paper appears to deal primarily with children, so is not relevant to this context.

---

***Name of paper:*** *The effect of de-individuation of the Internet Troller on Criminal Procedure implementation: An interview with a Hater*

**Paper Reference:** Bishop, J. (2013) International Journal of Cyber Criminology (IJCC), January – June 2013, Vol 7 (1): 28–48

**Brief summary of study characteristics**:

This paper provides an in depth interview with an Internet troller and discussion of the findings of this to provide a general framework for understanding these 'electronic message faults.' The interview with the troller makes it apparent that there are a number of similarities between the proposed anti-social personality disorder in DSM-V and flame trolling activities. An investigation into the application of the Criminal Procedure rules in United Kingdom finds a number of inconsistencies in the way the rules are followed, which it appears are causing injustices in the application of Internet trolling laws. Provides interesting context, but shuld not be relied upon as it makes quite bold claims on the basis of one interview.

# Appendix B: Definitions of online abuse identified within the literature

### Cyber-bullying

- An aggressive act or behaviour that is carried out using electronic means by a group or an individual repeatedly and over time against a victim who cannot easily defend him or herself (Myers and Cowie, 2017);
- Incidents which occur in a narrow range of circumstances—usually involving young people and educational contexts (Jane, 2015);
- Repeated verbal or psychological harassment carried out by an individual or a group against others through online services and mobile phones (European Parliament, 2016)

### Cyber-violence

- Behaviour by an actor which takes place online and which is hostile and aggressive, and which may also be offensive, indecent, obscene, or of a menacing character (Owens, 2016)

### Online antagonism

- A serious violation of a person's dignity, sense of security and freedom, as well as a violation of the private sphere (Enarsson and Naarttijärvi, 2016);

### Trolling

- To post deliberately inflammatory papers on a social media forum (Maltby, 2016);
- Deliberately attacking others online, typically for amusement (Hardaker, 2015);
- The online posting of deliberately inflammatory or off-topic material with the aim of provoking textual responses and/or emotional reaction (Jane, 2015);
- An attempt to argue with and upset people by posting inflammatory and malicious messages (Maltby, 2017);
- The act of deliberately posting inflammatory or confusing messages on the Internet in order to provoke a vehement response from a group of users' (Shaw, 2013);
- The posting of hateful comments by a person or group along with the more aggressive, premeditated and prepared hate movements undertaken by groups of people (Bratu, 2017);
- A range of antisocial online behaviours that aim at disrupting the normal operation of online social networks and media (Tsantarliotis et. Al; 2016);
- The targeting of defamatory and antagonistic messages towards users of social media (Williams and Pearson, 2016);
- A distinct new form of antisocial behaviour online (March, 2017);
- Flaming, griefing, swearing, or personal attacks, including behavior outside the acceptable bounds defined by several community guidelines for discussion forums (Cheng et. al, 2017);
- Online harassment against strangers, with the intentions of causing disruption and conflict for entertainment (Seigfried-Spellar, 2017);
- A form of behaviour through which a participant in a discussion forum deliberately attempts to provoke other participants into angry reactions, thus disrupting communication on the forum and potentially steering it away from its original topic (Hopkinson, 2013);

- Practice of behaving in a deceptive, destructive, or disruptive manner in a social setting on the Internet with no apparent instrumental purpose (Jonason et. al, 2014);
- Specific example of deviant and antisocial online behavior in which the deviant user acts provocatively and outside of normative expectations within a particular community; trolls seek to elicit responses from the community and act repeatedly and intentionally to cause disruption or trigger conflict among community members (Fichman and San-Filippo, 2015)

**Cyber-stalking**

- The repeated pursuit of an individual using electronic or Internet-capable devices…involves repeated unwanted communication, sexual advances or requests, and threats of violence. It also includes surveillance of a victim's location through diverse technologies (Henry and Powell, 2016)

**Cyber-racism**

- Racism that manifests in this online world…includes words, images and symbols posted on social media services, online games, forums, messaging services and dedicated 'hate sites'. Cyber-racism includes a wide spectrum of conduct in terms of seriousness and specificity, ranging from, for example, racist material disguised as 'humour' to direct threats and incitements to violence targeting specific individuals or groups on the basis of race (Mason, 2017);

**Flame-trolling/flaming**

- Heated online communications involving invective, insults, negative affect, and so on (Jane, 2015);
- Deliberately using emotionally charged or contrarian statements to engender a response (Cook et. al, 2017)

**Gender trolling**

- Online abuse targeted against women often with threats and/or fantasies of sexual violence (Jane, 2014)