Digital wildfires: hyper-connectivity, havoc and a global ethos to govern social media

Helena Webb
Marina Jirotka
University of Oxford, Department of Computer Science
Oxford, United Kingdom
helena.webb@cs.ox.ac.uk
marina.jirotka.cs.ox.ac.uk

Bernd Carsten Stahl
De Montfort University, Department of Informatics
Leicester, United Kingdom
bstahl@dmu.ac.uk

William Housley
Adam Edwards
Matthew Williams
Cardiff University, School of Social Sciences
Cardiff, United Kingdom
housleyw@cardiff.ac.uk
edwardsa2@cardiff.ac.uk
williamsm7@cardiff.ac.uk

Rob Procter
University of Warwick, Department of Computer Science
Coventry, United Kingdom
rob.procter@warwick.ac.uk

Omer Rana
Pete Burnap
Cardiff University, School of Computer Science and Informatics
Cardiff, United Kingdom
ranaof@cardiff.ac.uk
p.burnap@cs.cardiff.ac.uk

ABSTRACT
The last 5-10 years have seen a massive rise in the popularity of social media platforms such as Twitter, Facebook, Tumblr etc. These platforms enable users to post and share their own content instantly, meaning that material can be seen by multiple others in a short period of time. The growing use of social media has been accompanied by concerns that these platforms enable the rapid and global spread of harmful content. A report by the World Economic Forum puts forward the global risk factor of 'digital wildfires' – social media events in which provocative content spreads rapidly and broadly, causing significant harm. This provocative content may take the form of rumour, hate speech or inflammatory messages etc. and the harms caused may affect individuals, groups, organisations or populations. In this paper we draw on the World Economic Forum report to ask a central question: does the risk of digital wildfires necessitate new forms of social media governance? We discuss the results of a scoping exercise that examined this central question. Focusing on the UK context, we present short case studies of digital wildfire scenarios and describe four key mechanisms that currently govern social media content. As these mechanisms tend to be retrospective and individual in focus, it is possible that further governance practices could be introduced to deal with the propagation of content proactively and as a form of collective behaviour. However ethical concerns arise over any restrictions to freedom of speech brought about by further governance. Empirical investigation of social media practices and perspectives is needed before it is possible to determine whether new governance practices are necessary or ethically justifiable.

ETHICOMP, September 2015, De Montfort University, Leicester, United Kingdom.

Categories and Subject Descriptors
K.4 [**Computers and Society**]: Public and Policy Issues – *abuse and crime involving computers, ethics, regulation.*

General Terms
Management, Human Factors, Legal Aspects.

Keywords
Social media, governance, responsible research and innovation.

1. INTRODUCTION
The last 5-10 years have seen a massive rise in the popularity and use of social media platforms such as Twitter, Facebook, Instagram, Snapchat and Tumblr etc. A 2014 report prepared by the UK's independent regulator OFcom [1] found that over 80% of British adults go online regularly and that 66% of these have a social media profile. Social media platforms enable users to post their own content – news, opinions, images etc. – which is then available to be seen instantly either publicly or by audiences selected by the user. Most of these platforms also have functions that allow users to forward some of the content they see, through shares or retweets etc. This content therefore has the capacity to

be seen by multiple others across the globe in a short period of time.

This rapid spread of content via social media can offer undoubted societal benefits, such as the promotion of social cohesion through solidarity messages and humanitarian campaigns [2]. However as social media platforms have grown significantly in popularity, concerns have also risen over their capacity to enable the rapid spread of harmful content. Reports of 'cyber-bullying', harassment and 'shaming' on social media have become commonplace in popular media [3], whilst governments and other institutions have blamed platforms such as Twitter and Facebook for enabling the spread of false rumours [4] and inciting violence [5] during times of tension. These concerns have led to calls for more effective regulation of digital social spaces [6] – for instance through the criminalisation or restriction of certain kinds of user content. Inevitably however these calls are contradicted by other arguments that position the internet as a medium that supports and encourages freedom of speech and therefore label any increased regulation as unethical [7].

In this paper we take up these contemporary concerns over the propagation of content on social media and the appropriate governance of digital social spaces. We draw on a 2013 report by the World Economic Forum (WEF) [8], which describes the global risk factor of 'digital wildfires' – social media events in which provocative content of some kind spreads broadly and rapidly, causing significant harm. We discuss the WEF's report further in Section 2 and highlight a central question arising from it: does the risk of digital wildfires necessitate new forms of social media governance? In Section 3 we present the findings of a scoping exercise conducted to begin addressing this central question. Focusing on the UK context we present three short Case Studies of digital wildfire scenarios and then characterise the four key governance mechanisms relevant to the regulation of these scenarios. We identify gaps within current governance and in Section 4 suggest potential further practices that might be adopted to overcome them. We highlight ethical issues surrounding the adoption and justification of any new governance mechanisms. We also argue that empirical research is necessary to analyse the real time propagation of content on social media and investigate the practical experiences and perspectives of various stakeholders in the governance of digital social spaces. This empirical work will be taken up by the paper authors in further project work.

2. BACKGROUND

2.1 Social media and digital wildfires

In February 2013 the World Economic Forum published the report "Digital wildfires in a hyperconnected world" [see 8]. As part of the Global Risks series, the report describes the popular use of social media platforms as a serious threat to international security and societal well-being. Social media platforms enable information and misinformation to spread rapidly and reach huge audiences, so where this content is in some way provocative – for instance taking the form of rumour or hate speech, or containing politically or socially inflammatory messages – it can 'wreak havoc in the real world'. The report conceptualises these risks as 'digital wildfires': social media events in which provocative content of some kind spreads broadly and rapidly, causing significant harm.

The WEF report gives examples of potential digital wildfire scenarios. It describes how the spread of misinformation can cause harm because it has negative consequences before there is an opportunity to correct it. For instance the spread of unverified content can damage the reputation of an individual – as in the false naming in 2012 of a senior UK politician in connection to allegations of child abuse (see Case Study 1). It can also undermine the standing of commercial companies, organisations, or institutions - such as in false reports of British Army failures in Iraq in 2009. Furthermore it can undermine social cohesion, for instance by causing panic over apparent incidents of disease outbreaks and security threats or by reinforcing the 'groupthink' of individuals who position themselves in networks separate from the rest of society.

2.2 Digital wildfires and the governance of social media

The WEF report describes digital wildfires as arising from the 'misuse of an open and easily accessible system'. Social media platforms are widely and freely available to many users across the world and place relatively few constraints against provocative content in the form of an unverified rumour, inflammatory message etc. Given the negative consequences that this spread of provocative content can cause, the report asks whether digital wildfires can be prevented through effective governance. It notes that legal restrictions on freedom of speech are technically difficult to achieve internationally and ethically difficult to justify. Instead it argues that as digital social spaces continue to evolve, there is scope for the development of a 'global digital ethos' in which generators and consumers of social media adopt *responsible practices*. The development and promotion of this ethos are challenges that remain to be undertaken.

2.3 New practices towards a global ethos to govern social media?

The World Economic Forum's description of digital wildfires provides a useful means to conceptualise the risks posed by the propagation of provocative content on social media. Digital wildfires can be understood as fast paced phenomena involving a form of collective behaviour through the spread of content by multiple users. They can result in significant harms and present challenges to the effective and ethical governance of digital social spaces. If we accept digital wildfires as a global risk factor, we are led to examine the role of governance in regulating the 'havoc' they can cause and the potential for a global ethos promoting digital responsibility. Therefore the WEF report prompts a central question: does the risk of digital wildfires necessitate new forms of social media governance?

The remainder of this paper reports on a scoping exercise designed to begin answering this central question. Focusing on the UK context, we reviewed current social media governance relevant to digital wildfire scenarios. Through a series of case studies and the examination of relevant literature and resources, we identified four key governance mechanisms: legal governance, social media governance, institutional governance and user self-governance. We then identified the characteristics of these mechanisms and highlighted a number of gaps in their capacity to deal with digital wildfire scenarios. Whilst it may be possible to introduce further governance practices to fill these gaps, our scoping exercise reveals the need for further empirical investigation to determine whether new mechanisms are in fact necessary or ethically justifiable.

3. SCOPING THE CURRENT GOVERNANCE OF SOCIAL MEDIA

3.1 Rationale and questions for scoping exercise

The scoping exercise was conducted as part of an ongoing research project on the responsible governance of social media (see Section 5.3.2). We drew on the World Economic Forum report to pose a central question: does the risk of digital wildfires necessitate new forms of social media governance? In order to address this question we determined that it was also necessary to consider further questions:

- What governance mechanisms currently exist relevant to digital wildfires?
- How do current governance mechanisms map on to potential digital wildfire scenarios?
- Are there any gaps in current governance mechanisms?
- Could any gaps in current governance be filled by new mechanisms?
- (How) can new governance mechanisms be ethically justified?

3.2 Conduct of scoping exercise

Governance practices, in particular legal frameworks, can vary across countries. In order to produce specific findings that could map directly on to particular digital wildfire scenarios, we decided to focus on social media governance in the UK (where the project is funded and based). We identified a number of social media events that could be categorised as potential digital wildfire scenarios. We conducted case studies of these scenarios to identify: the kind of provocative content propagated across social media; the governance mechanisms applied and their impact; and questions and debates arising over the appropriate regulation of the scenario. Three of the case studies are summarised in Section 4.2.

Through the case studies we identified four key mechanisms that seem to operate in relation to digital wildfire scenarios in the UK: legal governance, social media governance, institutional governance, and user self-governance. We examined each mechanism in turn through reference to news reports, institutional reports and reviews, websites and social media platform Terms of Use etc. We assessed the scope of these existing mechanisms and identified gaps in their capacity to prevent or manage digital wildfire scenarios. We then identified a range of further governance practices that could potentially overcome these gaps. As these further practices might be seen to limit freedom of speech, this then led us to highlight important ethical considerations surrounding the regulation of digital social spaces. Finally, we reflected on our findings in relation to the central question posed by the scoping exercise.

4. THE CURRENT GOVERNANCE OF SOCIAL MEDIA IN RELATION TO DIGITAL WILDFIRE SCENARIOS

4.1 Overview of findings

In this section we present the results of our scoping exercise and describe the current governance of social media in relation to potential digital wildfire scenarios. We begin with short summaries of three digital wildfire case studies. We then identify and discuss the characteristics of four key governance mechanisms: legal governance, social media governance, institutional governance, and user self-governance.

Our results indicate that legal governance, social media governance and institutional governance all tend to be retrospective in character; they deal with the kinds of provocative content associated with digital wildfires after it has spread and had an impact. They also tend to act on individual users rather than the multiple users who may be involved in a digital wildfire. By contrast user self-governance appears to have a real time element and may have the capacity to limit the spread of content posted by individuals or multiple users.

4.2 Case studies of digital wildfire scenarios

In the first stage of the scoping exercise we identified events meeting the criteria of digital wildfires: that is, they involved the rapid and broad spread of some kind of provocative content on social media which caused significant harm to an individual, group, organisation and/or population. We drew up case studies of these scenarios to identify the different mechanisms that were applied to regulate the digital wildfire.

Three of the case studies are summarised here. They have been chosen as they exemplify: 1) the kinds of content that may be involved in a digital wildfire; 2) the different kinds of governance mechanisms that may be drawn on to regulate a digital wildfire; and 3) current debates around the appropriate regulation of digital social spaces.

4.2.1 Case Study 1: Lord McAlpine

On 2nd November 2012 a BBC television programme broadcast a report on the sexual abuse of children in North Wales care homes during the 1990s [9]. It revealed that two of the care home victims had identified a "leading politician from the Thatcher years" as one of their abusers. The broadcast did not name the politician concerned but – alongside subsequent reports from other news media – provided enough information to enable many people to infer that it referred to Lord Alistair McAlpine. People began to name him on social media - including Sally Bercow, political activist and media personality with over 55 000 followers. She posted the tweet shown in Box 1.

In the week following the broadcast it became apparent that McAlpine had been wrongly implicated in the report [10]. The BBC issued an apology and subsequently paid McAlpine £185 000 in damages. Some Twitter users immediately issued apologies for naming him. McAlpine and his legal team considered reporting the Twitter messages naming him to the police and then announced they would sue users for libel [11]. Experts were hired to collate all relevant tweets: around 10 000 tweets were identified as potentially defamatory – 1 000 original tweets and 9,000 retweets.

Box 1. Tweet posted by Sally Bercow

Ultimately, users with fewer than 500 followers were asked to make a charitable donation in return for having cases against them dropped and McAlpine announced his attention to pursue libel actions against 'high profile' users with more than 500 followers [12]. Whilst out of court settlements were reached with a number of these high profile figures, Bercow maintained that her tweet was not defamatory and the case was taken to court. At the trial, Bercow's argument that her tweet constituted a 'random' thought was rejected and the judge found that her reference to 'innocent face' was insincere and ironical [13]. The case was formally settled in October 2013. Bercow apologised for her 'irresponsible use of Twitter' and agreed to pay McAlpine undisclosed damages and cover his costs. She then temporarily closed her Twitter account. The case attracted a great deal of attention in the UK and was referred to by McAlpine's lawyer as "the leading case in terms of internet responsibility" [14].

4.2.2 Case Study 2: Caroline Criado-Perez

Caroline Criado-Perez is a journalist and feminist activist who was involved in a successful and high profile campaign in spring 2013 to guarantee a place for female historical figures (in addition to Queen Elizabeth II) on banknotes produced by the Bank of England [15]. Following the campaign, Criado-Perez wrote an article in the *New Statesman* revealing that she had been receiving numerous rape threats via Twitter from multiple accounts [16]. She reproduced some of the content of the tweets in the article (without including the account handles of the users who sent them) – see Box 2. Criado-Perez reported the tweets to the police and strongly criticised Twitter for not doing enough to deal with the threatening messages and the users who posted them.

"this Perez one just needs a good smashing up the arse and she'll be fine"

"*Everyone jump on the rape train > @CCriadoPerez is conductor"; "Ain't no brakes where we're going"*

"Wouldn't mind tying this bitch to my stove. Hey sweetheart, give me a shout when you're ready to be put in your place"

Box 2: Examples of abusive tweets quoted by Caroline Criado-Perez

The article provoked a range of discussion over the appropriate ways to deal with online harassment [17]. Some argued that reporting abuse to the police or social media platforms was unnecessary as users could 'use their own voices' to shame others who harassed them. However Criado-Perez maintained that the police and Twitter needed to do far more to help victims of harassment. A petition started in July 2013 calling for Twitter to simplify and speed up its systems for reporting abuse received 40 000 signatures in its first week [18]. In August 2013 the head of Twitter UK apologised to Criado-Perez for the abuse she had received and pledged that the platform would do more to stop similar abuse occurring [19]. Twitter subsequently introduced a 'report tweet' function that enabled users to report abuse immediately rather than having to send a message through its Help Centre [20].

In January 2014 Isabella Sorley and John Nimmo pleaded guilty to sending menacing messages to Criado-Perez [21]. It was stated in court that Criado-Perez had received abusive messages from 86 Twitter accounts, including multiple accounts held by the two defendants. It was also reported that Criado-Perez had suffered life changing psychological effects from the abuse she had received. Both Sorley and Nimmo received prison sentences and were described by their defence lawyers as naïve in their use of social media, taken in by the attention they received when their abusive posts were retweeted, and unaware of the harms they had caused.

4.2.3 Case Study 3:2011 England riots

On August 6th 2011 a peaceful protest over the police shooting of a man in south London became violent [22]. Over the next few nights disorder and looting spread across London and other towns and cities in England. Social media platforms such as Twitter and Facebook were widely used during this period and were seen by the government and some other commentators to play a significant role in enabling the spread of rumour, incitement of violence and organisation of gang activity.

The riots resulted in over 3 000 criminal prosecutions and a number of these involved the use of social media. For instance, Perry Sutcliffe-Keenan [23] received a 4 year custodial sentence after pleading guilty to intentionally encouraging another to assist the commission of an indictable offence. On August 9th Sutcliffe-Keenan had used his Facebook account to invite his 400 followers to riot in the town of Warrington the following day. However, he deleted the page shortly after setting it up and subsequently described it as a joke. No riots occurred in the town but the page was reported to the police by some members of the public. The court was told that Sutcliffe-Keenan's actions had caused panic in the local area and placed a strain on police resources. In another example a 17 year-old youth [24] was banned from social media sites for 12 months and ordered to complete 120 hours of community service after admitting sending a menacing message that encouraged rioting. He had posted a Facebook message saying "I think we should start rioting, it's about time we stopped the authorities pushing us about and ruining this country. It's about time we stood up for ourselves for once. So come on rioters – get some. LOL." The court heard that some of the youth's followers who saw the message replied by calling him an 'idiot' for posting it and the youth had deleted it by the time the police arrived to talk to him about it. No riots took place in the area where the youth lived and he told the court that the post had been intended as a joke.

The England riots prompted a great deal of discussion about the impact social media messages can have on offline behaviours and how/whether this should be governed. On August 11th Prime

Minister David Cameron announced that the government would review the possibility of preventing suspected rioters sending messages online [25]. In response to criticism of the site, a Facebook spokeswoman confirmed that the platform removed 'credible threats of violence' as part of its monitoring process [26]. She also pointed to the positive role that Facebook played during this time of great tension by providing a means for users to let family and friends know they were safe. Subsequent research [27] has suggested that the impact of social media in escalating the riots was overestimated; BBM smart phone messaging was used to coordinate illegal activity far more than social media and the response of Twitter and Facebook users to the unfolding events was more anti than the pro the riots. Many individuals took to social media to send messages condemning the violence and used the platforms to coordinate 'clean up' operations after the riots had ended.

4.3 Key governance mechanisms relevant to digital wildfires

The collation of case studies of digital wildfires enabled us to identify four key governance mechanisms relevant to digital wildfires. The characteristics of these governance mechanisms are discussed in turn.

4.3.1 Legal governance

In July 2014 the UK House of Lords Select Committee on Communications published a review of Social Media and Criminal Offences [28]. This concluded that, with the exception of criminalising online behaviours associated with 'revenge porn', it was not necessary to introduce new laws to govern social media in England and Wales. Therefore, legal actions regarding social media draw on existing civil and criminal legal codes. These can pursue individuals who have posted certain kinds of provocative content – such as defamatory claims (Case Study 1), menacing or obscene messages (Case Study 2) incitements to violence (Case Study 3), threats of violence, and breaches of court orders. Punishments for breaking these laws take the form of fines/damages, community service and custodial sentences.

In a typical digital wildfire scenario, a relatively small number of potentially illegal posts are reported to the police/lawyers and an even smaller number of these are pursued in the courts. In Case Study 1 the vast majority of users reached out of court settlements with the lawyers representing Lord McAlpine. In Case Study 2 the police were unable to identify all the users who had posted menacing messages and some cases were dropped as pursuing them was deemed not to be in the public interest [29]. In Case Study 3 only a very small number of users who posted inflammatory content about the riots were reported to the police.

Legal actions deal with provocative social media content retrospectively, after it has been posted, spread and had an impact. Beyond the use of deterrent sentences, legal governance therefore has little capacity to prevent the spread of provocative content and digital wildfires. Rhetoric around legal governance has frequently emphasised the limitations of the law in dealing with mass postings on social media [30]. It has also emphasised the responsibility of individual users to behave appropriately on social media (Case Study 1) and understand the potential impacts of their actions (Case Study 2).

4.3.2 Social media governance

Although social media platforms differ in the precise ways that they govern user content and behaviour, social media governance typically centres on the application of Terms of Use agreements. Platforms such as Twitter, Facebook, Flickr, Instagram, Tumblr etc. require users to sign up for an account by providing some contact and/or identifying information and agreeing to follow specific Terms of Use regarding what they can and cannot post on the platform. The Terms of Use generally set out penalties for breaches in the form of deletion of posts and suspension or closure of accounts.

Automated processes can identify and block certain types of content, such as explicit threats of violence (Case Study 3) and images of child sexual exploitation. However most often platforms rely on other users to report breaches of the Terms of Use. In some cases social media companies may pass on information to the police or security services, although they can be reluctant to do so [31].

Social media platforms often promote user self-governance. In addition to being able to report others, platforms typically have privacy and blocking functions so that users can control who has access to their posts. Certain features on a platform can encourage trust amongst users. For instance the use of real names and/or the addition of demographic information can help users to feel they 'know' each other. Users may also have the option to rate, rank, 'like' or 'favourite' others' posts to indicate that they – and by extension the user that posted them – are creditworthy. Similarly, users can draw on information about how many friends, followers etc. a poster has or how many posts they have made to draw conclusions about that poster's trustworthiness. Finally, some of the large social media service providers have taken part in awareness and education campaigns to promote responsible user behaviour [32].

The governance mechanisms of social media platforms are still evolving and changes are made on a regular basis. Twitter brought in significant changes to its reporting process following the abuse of Criado-Perez (Case Study 2) and has introduced further steps to tackle 'trolls' in 2015 following criticism from its own CEO [33]. However Twitter, like other social media platforms, is underpinned by the principle of freedom of speech and explicitly states that it upholds the right for users to post inflammatory content [34]. Sally Bercow's tweet in Case Study 1, although defamatory, did not breach Twitter's Terms of Use and the posts in Case Study 3 were not treated (at that time) by Facebook as credible threats of violence.

As with legal governance, the governance mechanisms within social media platforms focus on dealing with individual users and posts. Therefore they lack capacity to deal with the multiple posters involved in a digital wildfire scenario. Automated processes can to prevent the posting and reposting of certain kinds of content but most breaches are dealt with retrospectively and rely on user reports. As reporting can be a slow process, provocative posts can be often be seen and shared repeatedly – potentially causing significant harm - for a considerable period before they are acted on.

4.3.3 Institutional governance

As social media sites have grown in prevalence and popularity, organisations of various kinds have begun to institute policies to

govern appropriate content and user behaviour relevant to the particular institution. For example various employers require their employees to follow policies that outline what can and cannot be posted in official and personal accounts [35]. Typically, these place constraints on the posting of (negative) information about the employer organisation and can also extend to penalising users who undermine the organisation by behaving inappropriately – for instance by posting racist comments. Guidance to jurors in the UK now incorporates the use of social media [36] and many schools set out social media protocols to be followed by staff, students and parents [37]. Institutional governance appears to have some capacity to deal with the kinds of provocative content associated with digital wildfires as social media policies are likely to sanction certain kinds of unverified and inflammatory content. But once again this form of governance tends to be retrospective in focus and acts on individual users and posts after content has been spread.

4.3.4 Social media user self-governance

Users can undertake a number of actions that function to govern social media content. Where applicable they can report posts to the police or social media platform (Case Studies 2 and 3) or pursue other users through civil law (Case Study 1). They can set up privacy settings etc. to monitor who has access to their posts. They can delete or alter their own posts (Case Study 3) and even suspend their accounts (Case Study 1) where appropriate.

Users can also challenge content posted by others. For instance they might label a post as misleading or inappropriate. In Case Study 1 some of Bercow's followers urged her to remove her defamatory tweet and apologise for it before the trial, whilst work conducted on the 2011 riots found that users were able to successfully challenge and limit the spread of unverified rumours [38]. An alternative kind of challenge is to mock the poster in order to minimise the value of a post. For instance in Case Study 3 some followers labelled the youth an 'idiot'. Taken further, users also sometimes seek to 'shame' users for posting inappropriate content. This can be done in a variety of ways and includes: encouraging others to criticise a user; finding and spreading identifying details of the user; and passing on the user's posts to monitoring sites such as 'Yes, you're racist' or 'Racists getting fired'. Shaming can be highly effective in the sense that it can lead to users leaving the social media platform or losing their job etc. but it does raise ethical concerns over whether the harm it inflicts is justified by the harm caused in the offending post [39].

Finally, ignoring provocative posts and users has long been advocated as a way to deal with inappropriate content [40]. It stops content being spread and deprives users of the attention they are seen to crave. However since many social media posts have a very wide reach, it is perhaps unlikely that a large number of users will all ignore a provocative post. Furthermore some victims of online harassment (Case Study 2) argue that it is important to fight back against provocative posts rather than letting them pass without comment.

Self-governance practices appear to have some prospective characteristics. They may be able to counter the provocative content associated with digital wildfires in real time – for instance by challenging and correcting misinformation or preventing the spread of posts. Exactly how these practices play out during digital wildfires is a question that requires empirical investigation.

5. DISCUSSION

In this section we discuss the implications of the results of our scoping work. We describe gaps in current governance relating to digital wildfires and suggest further governance mechanisms that may overcome these gaps. We highlight key ethical questions regarding the introduction of any further governance practices and conclude that more empirical research is necessary to address our central question – does the risk of digital wildfires necessitate new forms of social media governance?

5.1 Current governance related to digital wildfires

5.1.1 Characteristics

We identified four current governance mechanisms relevant to digital wildfires: legal governance, social media governance, institutional governance and user self-governance. These mechanisms differ in the kinds of content they treat as inappropriate and in the kinds of sanctions they apply but all map on to digital wildfire scenarios to some extent.

Legal, social media and institutional governance mechanisms tend to be retrospective in focus as they deal with content after it has been posted. They typically apply sanctions to individual users. By contrast self-governance mechanisms have a real time element and may limit or prevent the spread of some posts.

Rhetoric surrounding these various mechanisms shares an emphasis on the importance of responsible user behaviour and can be seen to reflect the interest of the World Economic Forum in the development of a digital ethos that moves beyond legal regulation.

5.1.2 Gaps in current governance

None of the four governance mechanisms deal with digital wildfires as a specific phenomenon so it is inevitable that gaps in current governance arise. A key gap concerns the capacity for governance practices to act on multiple users rather than individuals. As described by the World Economic Forum, digital wildfires can be understood as involving a form of collective behaviour through the cumulative spread of content by multiple users. Legal, social media and institutional governance procedures focus on individual users and/or posts and therefore lack the capacity to deal with this characteristic. In addition, as these mechanisms – apart from the use of automated processes by social media platforms to block some kinds of content – deal with content retrospectively they do not have the capacity to prevent or limit the impact of digital wildfires in real time.

5.2 Potential further governance mechanisms

5.2.1 Types of mechanisms

It is possible that further governance structures could be introduced to map more directly onto the characteristics of digital wildfires and overcome some of the gaps noted above. This could include:

- Technical mechanisms to counteract the rapid spread of social media content. For instance the creation of a waiting time for retweets that could be linked to activity around a post or user. This would be comparable in principle to measures that slow down automatic trading when markets behave erratically.

- Further support for self-governance mechanisms that challenge and slow down the spread of provocative content. For instance the provision of visible esteem to individuals who intervened in the early stages of a digital wildfire in order to ensure the appropriate spread of content. Alternatively, the provision of a 'lie' button to indicate that the content of a post is not creditworthy or an 'ignore' button that users can activate to recommend that others do not respond to a post.
- Automated content analysis of posts to identify potentially defamatory, misleading, offensive etc. content. This could then trigger a warning to users recommending review of the post before submission.

5.2.2 Justification of governance

The alternative governance mechanisms suggested above are designed to limit the development and spread of digital wildfires and reduce the impact they can have. This is based on the assumption that digital wildfires are harmful and need to be limited. Insights from computer ethics and responsible research and innovation [41] illustrate the importance of ethical justifications for governance and in the case of digital wildfires this is not straightforward. Questions of harm and truth are central. Preventing the spread of provocative content can be beneficial but some mechanisms may produce more harm than the content itself. For instance, preventing or delaying the posting of content could be seen as a significant barrier to freedom of speech – and this in turn, as the World Economic Forum report acknowledges, can have very negative consequences. In addition the increasing prevalence of social media 'shaming' of posters can appear out of proportion to the harm done in an offending post. In any case how can the truthfulness or potential harmfulness of a post be established – and by whom? Wildfires that are based on truthful content may well be desirable – even if the content is provocative in other ways. Any consideration of governance mechanisms that limit digital wildfires needs to balance considerations of freedom of speech with issues concerning the avoidance of harm. This is in part a normative question but is also one that can be informed by empirical insights into how provocative content spreads on social media, the harms it causes and the capacity for existing governance mechanisms to deal with it.

5.3 Further questions

5.3.1 Need for empirical research

The results of our scoping exercise highlight the existence and characteristics of four key governance mechanisms operating in the UK context. We have shown that gaps in governance exist and that further governance practices may be possible but that these require careful ethical examination.

However this scoping exercise alone cannot answer the central question regarding the regulation of digital social spaces in the context of digital wildfires. Further questions emerge from our work which require empirical investigation. How do existing governance mechanisms operate in real time in digital wildfire scenarios? In particular, what role does self-governance play in limiting and halting the spread of provocative content? Furthermore, what kinds of harm do digital wildfires inflict on different individuals, groups, organisations and populations? Are these harms serious enough to support arguments for new mechanisms that will potentially limit freedom of speech? A better empirical understanding of digital wildfires is required to determine whether new governance mechanisms are necessary and justified, and what forms they might take. This important empirical work is taken up by the authors in our ongoing project – "Digital wildfire: (Mis)information flows, propagation and responsible governance."

5.3.2 The "Digital Wildfire" project

The "Digital Wildfire: (Mis)information flows, propagation and responsible governance" project [42] is an interdisciplinary study led by the University of Oxford in collaboration with the Universities of Cardiff, de Montfort and Warwick. The overall aim of the project is to build an empirically grounded methodology for the study and advancement of the responsible governance of social media in the context of digital wildfires. The scoping work discussed in this paper forms part of a review of existing governance mechanisms which will inform the empirical activities of the study. The empirical work will take 3 forms: 1) Case studies of 4 digital wildfires. We will collect digital media datasets and combine computational analysis with qualitative analysis to examine information flows during digital wildfires and the occurrence of self-governing behaviour, such as counter speech to combat rumour and antagonistic content. 2) We will conduct a series of online questionnaires to seek the informed opinion of various experts regarding the appropriate regulation of digital social media and digital wildfires. 3) We will conduct interviews and observations at various sites (such as social media platforms, police organisations, civil rights groups) to investigate and understand how stakeholders respond to instances where the digital spread of provocative content may create situations of offline tension, conflict or disturbance.

The results of the scoping and empirical work will be drawn on to produce an ethical security map. This will be a practical tool to help different users navigate through social media policy and aid decision making. Other project outputs include the development a training module on digital maturity and resilience for use in secondary schools and the production of artwork to promote a creative understanding of digital wildfires amongst a broad range of audiences.

6. CONCLUSION

In this paper we have drawn on the concept of digital wildfires – social media events in which provocative content spreads broadly and rapidly, causing significant harm – and reported on a scoping exercise conducted to investigate a central question: does the risk of digital wildfires necessitate new forms of social media governance? We have described and discussed existing governance mechanisms relevant to digital wildfires in the UK context and identified a number of gaps in current governance. We have highlighted opportunities for further governance practices that could overcome these gaps by prospectively preventing and limiting the spread of provocative content. We have also highlighted ethical concerns around the introduction of any new governance practices that might limit freedom of speech. The question of whether new governance approaches are necessary to regulate digital wildfires requires further investigation; we have demonstrated the need for empirical research that analyses the real time propagation of provocative content on social media and investigates practical issues and perspectives regarding its governance.

7. ACKNOWLEDGMENTS

The "Digital Wildfire: (Mis)information flows, propagation and responsible governance" project is funded by the Economic and Social Research Council. Project reference ES/L013398/1.

8. REFERENCES

[1] Ofcom. 2014 *Internet Citizens Report 2014* (Nov. 2014), DOI= http://stakeholders.ofcom.org.uk/binaries/research/telecoms-research/Internet_Citizens_Report_14.pdf

[2] Rotman, R., Vieweg, S., Yardi, S., Chi, E., Preece, J., Shneiderman, B., Pirolli, P. and Glaisyer. T. 2011. From slacktivism to activism: participatory culture in the age of social media. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '11). ACM, New York, NY, USA, 819-822. DOI=http://doi.acm.org/10.1145/1979742.197954.

[3] Ronson. J. 2015. *So You've Been Publicly Shamed*. (March 2015) Riverhead Books, London. 277 pages.

[4] Luckerson, V. 2014. Fear, misinformation and social media complicate ebola flight. t*ime.com* (Oct. 2014). DOI= http://time.com/3479254/ebola-social-media/

[5] Halliday, J. (2011) David Cameron considers banning suspected rioters from social media. *theguardian.com* (Oct. 2011) DOI= http://www.theguardian.com/media/2011/aug/11/david-cameron-rioters-social-media

[6] All party parliamentary group against anti-Semitism. 2015. *Report of the all party parliamentary inquiry into anti-Semitism* (Feb. 2015). DOI= http://www.antisemitism.org.uk/wp-content/themes/PCAA/images/4189_PCAA_Antisemitism%20Report_spreads_v9%20REPRO-DPS_FOR%20WEB_v3.pdf

[7] Foxton, W. (2015) Criminalising online trolls is absurd even if what they say is vile. t*elegraph.co.uk* (Feb. 2015) DOI= http://www.telegraph.co.uk/technology/social-media/11401308/Criminalising-online-trolls-is-absurd-even-if-what-they-say-is-vile.html

[8] World Economic Forum, 2013 *Digital Wildfires in a hyperconnected world*. Global Risks Report. World Economic Forum (Feb. 2013), DOI= http://reports.weforum.org/global-risks-2013/risk-case-1/digital-wildfires-in-a-hyperconnected-world/

[9] Greenslade, R. 2014. Newsnight's McAlpine scandal. 13 days that brought down the BBC's chief. *theguardian.com* (Feb. 2014). DOI= http://www.theguardian.com/media/greenslade/2014/feb/19/newsnight-lord-mcalpine

[10] Sabbagh, D. and Deans, J. 2014. BBC to pay Lord McAlpine £185 000 damages after false child abuse allegations. *theguardian.com* (Nov. 2012) DOI= http://www.theguardian.com/politics/2012/nov/15/bbc-lord-mcalpine-compensation-newsnight

[11] Swinford, S. 2012. Lord McAlpine vows to take on the "Twittering fraternity" *telegraph.co.uk* (Nov. 2012) DOI= http://www.telegraph.co.uk/culture/tvandradio/bbc/9688845/Lord-McAlpine-vows-to-take-on-Twittering-fraternity.html

[12] Dowell, B, 2012. McAlpine libel. 20 Tweeters including Sally Bercow pursued for damages. *theguardian.com* (Nov 2012). DOI= http://www.theguardian.com/tv-and-radio/2012/nov/23/mcalpine-libel-bercow-monbiot-davies

[13] Lord McAlpine of West Green v Sally Bercow. 2013. EWHC 1342 (QB). DOI= https://www.judiciary.gov.uk/judgments/mcalpine-bercow-judgment-24052013/

[14] Press Association. 2013. Lord McAlpine libel row with Sally Bercow formally settled in high court. *theguardian.com* (Oct. 2013) DOI= http://www.theguardian.com/uk-news/2013/oct/22/lord-mcalpine-libel-row-sally-bercow

[15] Bell, M. 2013. Victory for women on banknotes campaign. *Sofeminine.com* (July 2013) DOI= http://www.sofeminine.co.uk/key-debates/victory-for-women-on-banknotes-campaign-jane-austen-to-be-on-10-note-s86135.html

[16] Criado-Perez, C. 2013. After the Jane Austen announcement I suffered rape threats for 48 hours, but I'm still confident the trolls won't win. *New Statesman online* (July 2013). DOI= http://www.newstatesman.com/media/2013/07/after-jane-austen-announcement-i-suffered-rape-threats-48-hours-im-still-confident-tro

[17] Gold, T. 2013. How do we tackle online rape threats? *theguardian.com* (July 2013) DOI= http://www.theguardian.com/commentisfree/2013/jul/28/how-to-tackle-online-rape-threats

[18] Miller, B. 2013. UK petition calls on Twitter to tackle abuse after Caroline Criado-Perez subjected to violent tweets. *abc.net* (July 2013). DOI= http://www.abc.net.au/news/2013-07-29/thousands-sign-petition-to-stop-abusive-tweets/4849780

[19] BBC news. 2013. Twitter's Tony Wang offers apology to abuse victims. *bbc.co.uk* (Aug. 2013). DOI= http://www.bbc.co.uk/news/uk-23559605

[20] Doshi, S. 2014. Building a safer twitter. *Blog.twitter.com* (Dec. 2014). DOI= https://blog.twitter.com/2014/building-a-safer-twitter

[21] Cockerell, J. 2014. Twitter 'trolls' Isabella Sorley and John Nimmo jailed for abusing feminist campaigner Caroline Criado-Perez. *independent.org.uk* (Jan 2014). DOI= http://www.independent.co.uk/news/uk/crime/twitter-trolls-isabella-sorley-and-john-nimmo-jailed-for-abusing-feminist-campaigner-caroline-criadoperez-9083829.html

[22] Gentleman, A. 2011. London riots: social media helped gangs orchestrate looting, says MP. *theguardian.com* (Aug 2011). DOI= http://www.theguardian.com/uk/2011/aug/11/riots-social-media-gang-culture

[23] BBC News. 2011. England riots: two jailed for using Facebook to incite disorder. *bbc.co.uk* (Aug. 2011). DOI= http://www.bbc.co.uk/news/uk-england-manchester-14551582

[24] Miller. L. 2011. UK riots: 17 year old banned from using social networking sites for Facebook message. *Mirror.co.uk* (Aug. 2011). DOI= http://www.mirror.co.uk/news/technology-

science/technology/uk-riots-17-year-old-banned-from-social-185122

[25] Halliday, J. (2011) David Cameron considers banning suspected rioters from social media. *theguardian.com* (Oct. 2011) DOI= http://www.theguardian.com/media/2011/aug/11/david-cameron-rioters-social-media

[26] Halliday, J. and Garside, J. 2011. Rioting leads for Cameron to call for social media clampdown. *theguardian.com* (Aug. 2011). DOI= http://www.theguardian.com/uk/2011/aug/11/cameron-call-social-media-clampdown

[27] Lewis, P., Newburn, T., Taylor, M., Mcgillivary, C., Greenhill, A., Frayman, H. and Proctor, R. 2011. *Reading the Riots: investigating England's summer of disorder.* The London School of Economics and Political Science and The Guardian, London, UK. DOI= http://eprints.lse.ac.uk/46297/1/Reading%20the%20riots%28published%29.pdf

[28] House of Lords. 2014. *Social Media and Criminal Offences: 1st report of Session 2014-2015*. London. The Stationery Office Limited. DOI=: http://www.publications.parliament.uk/pa/ld201415/ldselect/ldcomuni/37/3702.htm

[29] Smith, J. 2013. Two to be charged with threatening tweets to campaigner who called for a woman to be on bank notes. *Mailonline* (Dec. 2013). DOI= http://www.dailymail.co.uk/news/article-2524784/Two-charged-threatening-tweets-woman-banknotes-campaigner-Carolina-Criado-Perez.html

[30] House of Lords. 2014. *Social Media and Criminal Offences: 1st report of Session 2014-2015*. London. The Stationery Office Limited. DOI=: http://www.publications.parliament.uk/pa/ld201415/ldselect/ldcomuni/37/3702.htm

[31] Independent Reviewer of terrorism legislation. 2015. *A question of trust. Report of the legislator powers review*. June 2015. DOI= https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/434399/IPR-Report-Web-Accessible1.pdf

[32] House of Lords. 2014. *Social Media and Criminal Offences: 1st report of Session 2014-2015*. London. The Stationery Office Limited. DOI=: http://www.publications.parliament.uk/pa/ld201415/ldselect/ldcomuni/37/3702.htm

[33] Tiku, N. and Newton, C. 2015. Twitter CEO: 'We suck at dealing with abuse'. *theverge.com* (Feb. 2015). DOI= http://www.theverge.com/2015/2/4/7982099/twitter-ceo-sent-memo-taking-personal-responsibility-for-the

[34] https://support.twitter.com/articles/20170133-offensive-content

[35] The Institute for Employment Studies. Workplaces and social networking. *ACAS*. DOI= http://www.acas.org.uk/media/pdf/d/6/1111_Workplaces_and_Social_Networking.pdf

[36] https://www.gov.uk/jury-service/discussing-the-trial

[37] Britland, M. 2012. Social media for schools: a guide to Facebook, Twitter and Pinterest. *theguardian.com* (July 2012). DOI= http://www.theguardian.com/teacher-network/2012/jul/26/social-media-teacher-guide

[38] Procter, P., Vis, F. and Voss, A. 2013. Reading the riots on Twitter: methodological innovation for the analysis of big data. *International Journal of Social Research Methodology*. 16,3, (Apr 2013) 197-214. DOI: 10.1080/13645579.2013.774172

[39] Ellis-Peterson, H. 2014. Mary Beard reveals she befriended twitter trolls following online abuse. *theguardian.com* (Aug. 2014). DOI= http://www.theguardian.com/books/2014/aug/27/mary-beard-befriends-twitter-trolls-online-abuse

[40] Nolan, G. 2013. Internet trolls thrive on attention – but please don't feed the animals. *The logical libertarian* (April 2013) DOI= http://logicallibertarian.com/tag/internet-troll/

[41] Carsten Stahl, B. 2012. Morality, Ethics and Reflection: A categorisation of normative research in IS research. *Journal of the Association for Information Systems*. *13*,8, (Aug 2012) 636–656. DOI= http://aisel.aisnet.org/jais/vol13/iss8/1/ ; Carsten Stahl, B. Eden, G. Jirotka, M. and Coeckelbergh, M. 2014. From Computer Ethics to Responsible Research and Innovation in ICT: The transition of reference discourses informing ethics-related research in information systems. *Information & Management*. 51, 6 (Sep 2014) 810-818. DOI= doi:10.1016/j.im.2014.01.001

[42] For more information see our project website www.digitalwildfire.org