Forecasting the governance of harmful social media communications: findings from the digital wildfire policy Delphi

Adam Edwards , Helena Webb , William Housley , Roser Beneito-Montagut , Rob Procter & Marina Jirotka

Published online: 21 Nov 2020.

Submit your article to this journal ☑

Article views: 60

View related articles ☑

View Crossmark data ☑

Routledge
Taylor & Francis Group

 OPEN ACCESS   Check for updates

Forecasting the governance of harmful social media communications: findings from the digital wildfire policy Delphi

Adam Edwards [a], Helena Webb [b], William Housley [a], Roser Beneito-Montagut [a], Rob Procter [c] and Marina Jirotka [b]

aSchool of Social Sciences, Cardiff University, Cardiff, UK; bDepartment of Computer Science, University of Oxford, Oxford, UK; cDepartment of Computer Science, University of Warwick, Coventry, UK

ABSTRACT

Social media exhibits the core characteristics of emergent technologies. It is disruptive of established ways of organising social relations, is evolving at an exponential pace and its effects, including the production of new 'goods' and 'bads', are highly uncertain. Interest in understanding these effects has intensified in the context of fears over so-called 'digital wildfire', a policy construct referring to rapid propagation of harmful communications, particularly those involving children and other vulnerable social groups but also those threatening the integrity of the political process in liberal democracies. Even so, proponents of social media are anxious to protect its potential for enhancing freedom of speech and revitalising civil society through the redistribution of editorial powers to shape public debate and facilitate the democratic scrutiny and oversight of elites. This article reports findings of the 'Digital Wildfire policy Delphi', which asked key informants to consider the political and technical feasibility of regulating harmful social media communications and to forecast likely scenarios for their prospective governance. Key forecasts are that forms of enforcement are limited, stimulating 'self-regulation' will become increasingly important but, more controversially, the likelihood is that harm to vulnerable groups will be 'accommodated' in liberal democracies as a price to be paid for the perceived political and economic benefits of unmoderated social media. The article concludes with conjectures about future directions in the policing of social media and their implications for shaping the emerging research agenda.

1. Introduction

The governance and regulation of social media communications is a contemporary issue of technical and political controversy. Social media platforms equip any citizen with access to the internet with the capacity to broadcast their own opinions globally and in real-time. This circumvents traditional editorial and censorial control over who can contribute to public debates. As such, social media communications have been celebrated as a progressive disruption of powers hitherto wielded by press barons, corporations and political parties (Edwards *et al*. 2013, Housley *et al*. 2014). However, this enthusiasm has been increasingly qualified by a concern with the harmful effects of unmoderated, 'user-generated content' (UGC). Such communications include, *inter alia,* defamation and reputational damage to individuals, generic abuse and campaigns targeted at entire social demographics

---

and propagation of rumour and 'fake' news (Procter *et al*. 2013, Webb *et al*. 2015, Zubiaga *et al*. 2018, Procter *et al*. 2019). Research also suggests that as online abuse increases, then so does abuse off-line (Beran and Li 2007, Juvonen and Gross 2008) and that vulnerable groups such as children, adolescents and ethnic minorities can be particularly targeted and affected.

The potential harms of social media communications and the challenges they pose to governance were encapsulated in a World Economic Forum report in 2013.[1] This describes a modern condition of hyperconnectivity, in which social media platforms enable individuals to communicate spontaneously and with multiple others. This hyperconnectivity is a global risk factor for 'digital wildfires' in which harmful or misleading content (whether intentional or unintentional) spreads virally and causes damage to the reputation and wellbeing of individuals, groups and communities, as well as potentially affecting markets and other institutions. According to the report, digital wildfires create 'havoc in the real world' and can generate social tension during critical events such as civil disturbances, health scares and natural disasters. Significantly, the viral properties of social media means that content can propagate and have negative impacts before other agencies – such as the police, news media etc. – have a chance to respond. False or misleading content can become embedded and acted on before it can be corrected and harmful content can damage reputations and wellbeing. Consequently, there is a great deal of interest in identifying governance solutions to prevent, manage and mitigate digital wildfire type events (Webb *et al*. 2016, Procter *et al*. 2019). This may involve traditional governance mechanisms associated with regulation and policing as well as actions by social media platforms and users themselves. The WEF report highlights that discussions of effective governance are also complicated by the desire to protect freedom of speech.

Since the WEF report's publication, many instances of malicious and rapidly spreading social media communications have occurred. These have involved, for instance, the unregulated use of advertising on platforms, such as Google's video-sharing service YouTube, to propagate politically extreme and sexually explicit content[2] or the use of micro-blogging platforms, like Twitter, to target individuals or entire social groups for abuse (Housley *et al*. 2017, Procter *et al*. 2019). The 2016 US Presidential election and its aftermath became a flashpoint for discussion of 'fake news' and ways that the persuasive properties of social media communications might be weaponised by commercial and state actors.[3] In addition, various crises and critical public debates – such as the COVID-19 pandemic – have been accompanied by conspiracy theories, claims and counter claims of 'fake news' on social media (Ahmed *et al*. 2020, Depoux *et al*. 2020).

Debates over policing and governance of social media have also grown more pressing in this period. The law enforcement response differs across national jurisdictions, with different countries taking different approaches to the question of whether social media posts should be legally actionable. In the UK, posts and users can be dealt with through civil and criminal law in some circumstances – although the threshold to warrant criminal prosecution is high and can be hard to attain (Crown Prosecution Service 2018). Notable cases include the libelling of Lord McAlpine as a paedophile[4], sexist abuse of the feminist campaigner Caroline Criado-Perez[5] and racist abuse of female politicians[6] – the latter cases leading to prison sentences for some of the perpetrators. However, the number of harmful social media communications dealt with this way is small and represents only a tiny of proportion of those that occur.[7] This has led to further discussions about what social media users should do to regulate their own and others' behaviours online, as well as scrutiny of the role of platforms in governing social media content (Procter *et al*. 2019). Whilst social media platforms are known to deploy censorship algorithms to deal with some forms of content – for instance, to identify and block images of child sexual abuse – they typically rely on reporting mechanisms in which users are required to highlight problematic posts, which may then be removed after a period of inspection (Parliament House of Lords 2019). Legally, platforms have historically been viewed as infrastructure providers similar to telecommunications companies and free from being held responsible for the content they host – in contrast to traditional media such as newspapers (Jarvis 2018). The large social media platforms – perhaps reflecting their USA origins – also tend

towards positions in favour of upholding freedom of speech and reluctance to remove content (McNair 2018). For instance, Twitter and Facebook have typically removed reported content in cases of clear illegality but sometimes refused to remove posts that, while highly inflammatory, may not meet legal thresholds of hate speech and do not breach their terms and conditions. However, public authorities and mainstream news organisations have increasingly argued this stance is no longer credible. In 2019, the UK government announced proposals to legislate for a statutory duty of care by social media firms (DCMS 2019). The COVID-19 crisis has led to an increased focus on harms caused by conspiracy theories and misleading content, resulting in public commitments from social media platforms to take steps such as labelling suspect content and elevating authoritative content.[8]

In the current landscape, digital wildfires are a regular occurrence and questions of governance remain unresolved. The debate continues: should further governance mechanisms be put in place and, if so, what forms should these mechanisms take – law enforcement, self-regulation or other? Public authorities in liberal polities are presented with a number of strategic dilemmas: the reduction of harms, balanced with protection of freedom of speech and upholding democratic participation; surveillance of real time communication for purposes of harm reduction, balanced with protection of citizens from intrusion. A general challenge for all with an interest in this debate is the constantly developing, changing and shifting affordances of social media platforms and manner of their use. This means that social media communications and their consequences are hard to predict and pre-empt.

Dilemmas generated by social media communications are of immediate concern to public authorities and are set to intensify. It is also clear that there is both great urgency and scope for academic research to contribute nuanced understandings to this debate, including identifying opportunities for governance. Furthermore, we cannot be content with retrospective analyses: we have an obligation to engage in forecasting given the uncertainty and severity of harms associated with this emergent technology. This article reports on a policy Delphi held to foster such forecasting of prospects for the governance of social media communications. This deliberative method is designed to elicit rival viewpoints that can be used as a resource for anticipating processes of social change, in turn informing policymaking (de Loe et al. 2016). Such forecasting does not seek to predict specific events but to identify a range of plausible scenarios that can provide defensible grounds for the anticipatory governance of uncertain social problems. This enables nuanced discussion that acknowledges the kinds of strategic dilemmas noted above and accommodates the changing properties of social problems, such as the goods and bads of emergent technologies, provoking debate over policy responses that address harmful consequences without compromising perceived benefits.

To this end, the Digital Wildfire policy Delphi conducted interdisciplinary analyses of malicious social media communications and their impact to determine whether risks presented by these communications necessitate new forms of social media governance. The remainder of this article elaborates the value of the policy Delphi as a deliberative method for investigating the challenges of governing social media communications. It discusses its application in the Digital Wildfire study and relates findings from this study to concepts of enforcement, self-regulation and accommodation found in the broader literature on governance. In this context, governance is conceptualised broadly to include traditional forms of regulation and policing as well as other mechanisms involving social media platforms and their users. The limitations of enforcement methods, such as criminal prosecution, licensing and 'disruption' of abusive communications, are contrasted with the prospects for non-enforcement methods, such as stimulating 'self-regulation' amongst social media users. Non-enforcement is also considered in terms of the likelihood of harms, even to vulnerable social groups, being 'accommodated' as an acceptable cost for the perceived benefits of these technologies.

2. Methodological approach: the policy Delphi

The policy Delphi is a type of deliberative method in social science concerned with identifying rival clusters of opinion about a problem amongst a panel of key informants. These informants are sampled purposively on the grounds of their experience and/or expertise regarding the problem in question rather than their representativeness of some or other population (Turoff 1970, de Loe *et al*. 2016). Given this analytical concern with problems not populations, the number of informants recruited is often small relative to sample surveys, especially where the policy problem is highly focussed, less exploratory and reliant upon specific expertise (de Loe *et al*. 2016, p. 82). This logistical case for limiting the sample of informants was initially made by Turoff (1970, p. 153), who argued panels would ordinarily entail a range of 10–50 panellists: large enough to identify clusters of pre-dominant, minority and even outlier opinions but small enough to facilitate panellists' deliberative engagement with each other's insights, especially where the problem is multi-faceted and not easily reduced. Turoff's original concern was to use this method to extend involvement in deliberation about policy issues beyond the more restrictive coteries of elite decision-makers routinely found in public administration committee structures. As such, the policy Delphi method was initially thought of as a means of disrupting received wisdom in order to generate rival insights, equipping decision-makers with a broader palate of ideas to better forecast and govern policy problems about which there is limited knowledge, and as a method for countering confirmation of bias and the allied reproduction of theory failures (Turoff 1970, p. 152). Recent innovations in methodological thought provide additional grounds for justifying this scale of purposive sampling. They are concerned with the importance of limiting, rather than extending, the sampling of informants to those whose expertise and/or experience are particularly relevant to the problem in question. Elsewhere in policing and security studies, such informants have been referred to as 'sentinels' (Lowe and Innes 2012), an evocative concept used to distinguish between informants whose insight is likely to prove especially significant, by virtue of their direct involvement in the problem in question, from those experiencing the problem second-hand. In this regard, and given the conceptual framework adopted in the Digital Wildfire research (see 2.1., below), this implied the purposive sampling of authorities responsible for taking action on abusive social media communications, specifically legal and educational actors along with the potentially disruptive influence of the social media platforms themselves.

A policy Delphi is also distinguished by its combination of: (i) a longitudinal research design, in which deliberation occurs amongst the same panellists over time and through a series of iterative rounds of debate that provide a means of both construct and respondent validation; and (ii) methods of data collection that are anonymous and asynchronous, such that panellists do not know each other's identity and respond to the panel co-ordinator remotely, for example, through email and/or on-line questionnaires. Such anonymous deliberation liberates respondents to express genuine views and particularly benefits those who are professionally or even (in the case of civil servants) constitutionally obliged to represent certain positions in open forums rather than expressing their own informed opinions and values. A fundamental advantage of this anonymity is the removal of overt social hierarchies and allied power relations that can privilege the voice of some whilst silencing others. Dialogue between panellists is facilitated by the co-ordinator who summarises the viewpoints of panellists expressed in each round and reports this back to the panel as a whole. Typically, 2–3 rounds of questionnaires and reports take place.

As a relatively undeveloped approach, there are various areas of methodological debate surrounding the policy Delphi and opportunities for cumulative learning across its user community (de Loe *et al*. 2016). With this in mind, we describe our methodology in detail by setting out the study's research *strategy,* including its conception of policing and regulation, *design* and *methods*.

2.1. Research strategy and conceptual framework

The iterative deliberation inherent in the policy Delphi problematises the relationship between theory and empirical observation throughout the research process. The strategy is 'adaptive' (Layder 1998), in that theoretical propositions are used to structure the iterative design of the policy Delphi but are subsequently adapted in relation to the emerging findings from each round of deliberation. This dialogue between theory and observation is elaborated in the discussion of findings from each round.

In the study, the policy construct of 'digital wildfires' was initially framed as a problem of governance and regulation, and how allied theories drove the research design. Through reference to research, more specifically, on the relationship between 'policing and regulation' (Gill 2002, Edwards and Gill 2002), a review was undertaken of key contemporary governance mechanisms for malicious social media communications. Focusing on England and Wales, the review revealed four mechanisms and theoretical propositions for subsequent investigation:

- *Legal governance*: Certain social media behaviours may be actionable under (criminal or civil) law. Legal actions deal with social media content retrospectively, after it has been posted, spread and had an impact. They effectively target only a small number of those involved in posting or sharing content that has caused harm. Beyond the use of deterrent sentences, legal governance, therefore, has little capacity to prevent the spread of digital wildfires in real time.
- *Platform governance*: As with legal governance, governance mechanisms within social media platforms focus on dealing with individual users and posts. Therefore, they lack capacity to deal with multiple posters involved in a digital wildfire scenario. Automated processes can prevent 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, harmful posts can often be seen and shared repeatedly – potentially causing significant harm – for a considerable period before they are acted on.
- *Institutional governance:* social media policies are increasingly used in workplaces, places of education, hospitals and other institutions. They serve to sanction posting of certain kinds of unverified and inflammatory content but are typically focused on preserving the reputation and integrity of the institution itself. Once again, this form of governance tends to be retrospective and acts on individual users and posts after content has been spread.
- *Self-governance:* users have the opportunity to monitor and regulate their own social media behaviours, and potentially do the same to others. For instance, they can challenge and correct posts containing misleading information, counter inflammatory opinions, or discourage the sharing of posts. This has the capacity to influence propagation of content in real time and therefore may be able to play a role in preventing and limiting digital wildfires, especially where countervailing points are made by reputable, authoritative, accounts. This has the advantage of reducing harms without stifling legitimate free speech.

Alongside these theoretical propositions, empirical research on social media interactions was also used to inform the focus of the policy Delphi. A key finding from analysis of sequential Twitter 'threads', referring to interactions between micro-bloggers retweeting, mentioning and responding to provocative posts, is their heterogeneity: they may agree, disagree, praise, criticise, echo and request further information in their responses (Housley *et al*. 2018, Procter *et al*. 2019). This interactional analysis emphasises the vitality and potential for self-regulation of malicious communications amongst 'producers'[9] of social media. Indeed, these findings suggest that users of social media are amongst the most potent governors of abusive communications, intervening to challenge and correct inflammatory content or unsubstantiated rumour (Housley *et al*. 2018).

This review of the literature and emerging empirical evidence about interactions on social media helped to shape the design of the Digital Wildfire policy Delphi and identified issues

that could stimulate deliberation amongst its panellists. In particular, the study drew upon an existing conceptual framework about the relationship between policing and regulation to better relate emerging research on social media communications to the domain of policing and society (Edwards and Gill 2002, Gill 2002). Whilst this framework was originally developed in relation to understanding the problem of policing offline 'transnational organised crime', its identification of a spectrum of enforcement and non-enforcement strategies for responding to security problems helps to clarify insights emerging from research into social media interactions and their relationship to wider questions of policing and regulation. A central claim of this framework is the inadequacy of criminal prosecution as a means of reducing harms associated with high volume offending and victimisation. This point is echoed in work noting the limited capacity for prosecutions to deal with the large number of social media posts and users involved in the propagation of communications already prohibited under legislation, such as the Malicious Communications Act 1988 in England and Wales (Webb *et al*. 2015). Of relevance to the focus on harmful social media communications, drawing on this framework, it is possible to identify the following types of interventions entailed in a broader concept of policing beyond that of criminal law enforcement:

(1) The use of *administrative penalties* – such as the $5bn fine levied in July 2019 on the social media company Facebook by the US Federal Trade Commission for violating the privacy of users' personal data – offers a significant contrast to the relatively high costs and limited impact of prosecution. Such penalties also shift the focus of regulation from users of social media platforms to the platforms themselves and their allied tech companies.

(2) *Disruption*, which is preventive and concerned with reducing the opportunities for harmful social media communications to occur. For instance, alterations to the capacity of users to post or receive harmful content, such as automatic blocking of content by platform censorship algorithms, time delays on sharing content coming from 'red flagged' accounts, and provision of 'report' or 'panic' buttons to flag particularly harmful posts. Another disruption is through software altering users' access to certain platforms during certain periods of time, for example, young peoples' access to social media platforms during the peak period for online victimisation, estimated as 22.00–24.00hrs.[10]

(3) *Accommodation* reflects the reality that regulatory problems invariably escape the capacity of responsible authorities to pursue policies of full enforcement and are thus, effectively, accommodated. Most obviously in the case of harmful social media, the sheer volume and velocity of communications that defy criminal prosecution and, self-evidently, the negligible deterrent effect of the threat of such prosecution.

(4) In the condition of *regulatory capture,* a more extreme variant of non-enforcement, regulators become captives of the regulated because of shared ideology and personnel and/or shared rewards and mutually perceived threats, ensuring non-enforcement or very limited government regulation. In the case of social media, regulators, especially in the United States in which many of the principal social media companies are head-quartered, share the libertarian ideals of these companies (Samples 2019).[11]

(5) Counterpoised both to variants of enforcement by responsible authorities and to variants of non-enforcement that effectively give up on harm reduction, are strategies of *self-regulation*. These can include cultivation of resilience, particularly amongst vulnerable groups, for example, through educational campaigns with school children about dangers of online communications. Self-regulation can also refer to collective (rather than individuated) self-policing, for example, through counter-speech in response to breaches of norms of communication amongst online communities.

2.2. Research propositions

It is argued this conceptual framework provides a series of resources for deliberation about harm reduction and freedom of speech in the governance of social media. In addition to arguments about the limits to enforcement through criminal prosecution, the framework suggests the following propositions:

- Self-regulation provides a means of tackling harmful social media communications without draconian restrictions on freedom of speech.
- Advances in machine learning can disrupt high volumes of abusive social media communications in near or real time through automated censorship.
- Non-enforcement will lead to the continued, high-volume, abuse of vulnerable groups on social media.
- Where limits to disruption and non-enforcement result in continued and high-volume abuse of vulnerable groups on social media, liberal governments might consider more draconian restrictions on freedom of speech.

2.3. Research design

The policy Delphi method follows a longitudinal design aimed at validating, challenging and adapting panellists' constructs through iterative rounds of structured debate. The Digital Wildfire policy Delphi was designed in three rounds:

The first round (R1) asked panellists open-ended questions about 'digital wildfire' as a meaningful construct, its principal characteristics and, mindful of the conceptual framework, what, if any, challenges it generates for policing and public protection, whether these challenges ought to be regulated and, if so, how.

The second round (R2) questionnaire was accompanied by a report on panellists' responses to the R1 questionnaire and asked panellists to rank their agreement or disagreement about the construct of digital wildfire and alternative concepts of 'harmful' social media communications. It also asked panellists to rank the technical and political feasibility of the various regulatory strategies identified in the first round. Respondents were asked to rank the feasibility of particular strategies according to a Likert scale from '1' (definitely feasible) through to '5' (definitely unfeasible). They were also invited to report the reasoning behind their choices in free text boxes.

The third round (R3) questionnaire, accompanied by a report on responses to R2, asked panellists to forecast which scenarios for regulating harmful social media communications they thought most likely, given views expressed in the second round about the technical and political feasibility of different regulatory strategies.

The panel was composed by recruiting informants from four sectors and allied sub-sectors of expertise and experience, which are summarised in Table 1. These sectors were identified through the governance review and allied conceptual framework described above and the sub-sectors

Table 1. Sectors and sub-sectors used in the Digital Wildfires policy Delphi.

Sectors	Sub-sectors
User/scientist	*Academics – policy, social science/sociology, computer science, ethics, criminology*, responsible research and innovation, cyber security
Social media platforms	*Members of social media/online platforms with responsibility for: technology and infrastructure and public policy*. Members of social media/online platforms who conduct research and are responsible for strategic research programmes.
Institutions	*Law enforcement, teachers with interest in e-safety, members of anti-hate organisations*, members of city councils, members of policy think tanks, members of privacy organisations.
Lawyers	*Lawyers with an interest in Internet policy and law, data handling, privacy, data regulation, freedom of speech, ethics*. Lawyers with an interest in new media and technology, risk and negligence, defamation, legal history, press regulation, medical law, commercial law.

highlighted in *italics* are those from which respondents were successfully recruited. To re-emphasise, the broader conceptualisation of policing and regulation provided by this framework acknowledges the continued centrality of arguments over the significance of criminal and civil law enforcement, hence the purposive sampling of lawyers involved in the prosecution of malicious social media communications and the online services that facilitate them. However, as also noted above, this framework extends the conceptual and empirical scope of research to encompass other kinds of actors with analytical insight into policing such communications through other regulatory strategies. Of particular importance here are the social media platforms themselves, given the increasing interest in automated censorship that could disrupt malicious communications proscribed by criminal law without suppressing free speech where this law is silent. Given the study's other theoretical propositions, key informants from institutions involved in non-enforcement responses to malicious social media communications, particularly those aimed at equipping vulnerable groups with a greater resilience (police crime prevention officers, school teachers, anti-racist campaigners), were purposively sampled. Finally, given the pace at which this policing challenge is evolving, the study sought the expertise of others in social and computational research communities involved in investigating attempts to regulate malicious social media communications.

The composition of the Digital Wildfire policy Delphi panel is also related to epistemological assumptions underpinning this method. To reiterate, policy Delphi studies are interested in deliberation about problems amongst purposively sampled informants whose anonymous insights are liable to be analytically significant by virtue of their direct expertise and experience of the problems in question. They are not designed to represent the experiences of populations, even the populations of analytically significant informants, such as the views of *all* social media platforms, schoolteachers, legal practitioners, or social and computational researchers concerned about malicious social media communications. As such, the external validity of policy Delphi studies is in their adaptation of conceptual insights into policy problems, not the generalisability of their findings. On the question of generalisability as a research goal, a particular strength of policy Delphi studies is their revelation of minority, even outlier, opinions that may not be representative of populations but can, nonetheless, be prescient sources of innovative thinking precisely because they don't reflect mainstream opinion. In summary, the 19 respondents who participated in R1 represent a third of the 56 potential respondents who were initially approached, which, as noted above, is within the range of 10–50 respondents conventionally recruited to policy Delphi panels (Turoff 1970, de Loe *et al.* 2016). Of the 19 respondents to R1, 17 participated in R2, an attrition rate of 11%, which is beneath the round-to-round attrition rate of 30% conventionally regarded as acceptable in Delphi research (Sumsion 1998), as was the 23% attrition rate between the second and final rounds, culminating in 13 panellists involved in R3, again, remaining within the 10–50 range acknowledged as acceptable. Figure 1 summarises the profile of panellists in this final round. There was a continuity of panellists across the three rounds and, as such, panellists in the final round had also participated in the previous two rounds of deliberation in reaching their forecasts about the governance of harmful social media communications.

2.4. Research methods

Given that the central objective of the policy Delphi is to elicit rival viewpoints about the policy problem in question and retain a concern with those differences of opinion, analysis focuses on the clustering of responses around a range of views, including outlier insights that may be sources of innovative thinking outside of mainstream wisdom. The aim is to understand how initially held views are entrenched or altered because of subsequent rounds of deliberation and how this clarifies the range of plausible scenarios about the policy issue in question (de Loe *et al.* 2016). Identifying the plurality of plausible scenarios is central to the distinction between specious prediction and informed forecasting about highly uncertain policy problems in conditions of accelerated social change, as epitomised by the problem of policing online harms.
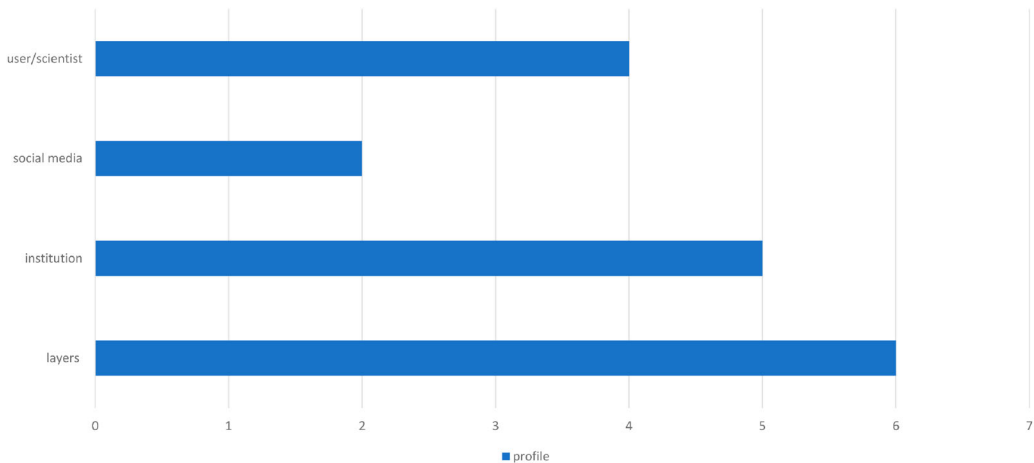
Figure 1. Profile of panellists in the final round of the Digital Wildfire policy Delphi.

Stacked bar charts have been used to demonstrate the breadth of opinion amongst panellists and quotations of free text responses contextualise the justifications for their informed judgements in response to questionnaires in each of the three rounds. This approach allowed the capture of both outliers and central tendencies of opinion about the technical and political feasibility of governing harmful social media communications and future prospects for governance of this problem. Findings from the Digital Wildfire policy Delphi are considered in terms of the sequential R1, R2 and R3.

3. Findings

The study entailed an evolving dialogue between theory and empirical observation throughout each of its rounds of deliberation. The key findings relate to:

- *construct validation* of the concept of 'digital wildfire' (considered in R1 and R2)
- *political and technical feasibility* of different regulatory strategies for governing social media communications (considered in R2 and R3)
- *future scenarios* for governing social media communications (considered in R3)

3.1. Construct validation

In R1, respondents were provided with a definition and some contextual examples of digital wildfires. They were then asked to write free text responses to the question: *Would/do you recognise digital wildfires and if so, what do you think are their main characteristics*?

A large majority of respondents across all sectors said they do recognise digital wildfires[12] and described their characteristics in terms of:

- *actions* – high volume and fast spread of content, similar to information cascades and speculative bubbles in markets. Posts made by users who are not direct witnesses of events;
- *content* – text and images that are negative and provocative, false or unverified information that is misleading, antagonistic, sexist, racist; and
- *consequences* – ignite debate, cause offence, are violent, are confusing when offline and online versions do not match, have a negative impact on society whether intentional or unintentional.

Within this overall recognition, however, several panellists pointed to limitations of the concept. They argued that the notion of digital wildfire is open to interpretation and can be hard to identify objectively whilst the consequences of a digital wildfire can be ambiguous:

It is impossible to tell a digital wildfire from genuine interest of a news story, gossip, moral panic, and even humour. (L6)[13]

I would urge caution against using the term 'digital wildfire' which implies that the dissemination of content is somehow 'out of control' … this language could be used to promote greater regulation of content on social media platforms in circumstances where it would be neither necessary nor desirable. (L4)

Information spreading from major [social media] is fallible; its validity or authenticity relies strongly on the reputation and trustworthiness of its authors, their institutions (example: a newspaper or news channel) or regulating body (community administrators). (U4)

Drawing upon R1 responses, R2 presented panellists with a series of Likert scale questions asking them to indicate their level of agreement with the following statements:

Q2.1.1.[14] 'Digital wildfire' is an ambiguous term that confuses more than it enlightens policy debates over the regulation of harmful social media communications.

Q2.1.2. For the purposes of regulating harmful social media communications, the term 'digital wildfire' is better replaced with references to specific offences (such as 'defamation', 'incitement', 'libel', 'menacing' and 'obscenity').

Q2.1.3. Social media communications should only be regarded as harmful if they can be unambiguously related to an existing offence (such as 'defamation', 'incitement', 'libel', 'menacing' and 'obscenity').

Figure 2 illustrates the overall clustering of responses. It shows that participants were divided in their opinion of the usefulness of digital wildfire as a concept. Their free text responses to the questions further revealed their ambivalence. Those in favour of the construct described how 'digital wildfire' can engage and enlighten various audiences. Those against highlighted that it can relate to very diverse types of communication that can be interpreted differently by audiences. There was a stronger clustering of opinion in agreement with replacing the construct with a focus on specific offences. However, some panellists cautioned against reducing the problem of harmful social media communications to issues of criminal law enforcement. They pointed out that 'harm' is also an ambiguous term and that not all forms of harm are (or should be) illegal.

Overall R1 and R2 clarified the strategic dilemma regarding harm reduction and freedom of speech in the governance of social media. One aspect of this concerns the use of more precisely defined constructs of problematic social media communications, those that are offences against criminal law, as the only grounds justifying regulation. These more tightly defined constructs of the problem limit scope for intervention and can act as a bulwark against looser categories of
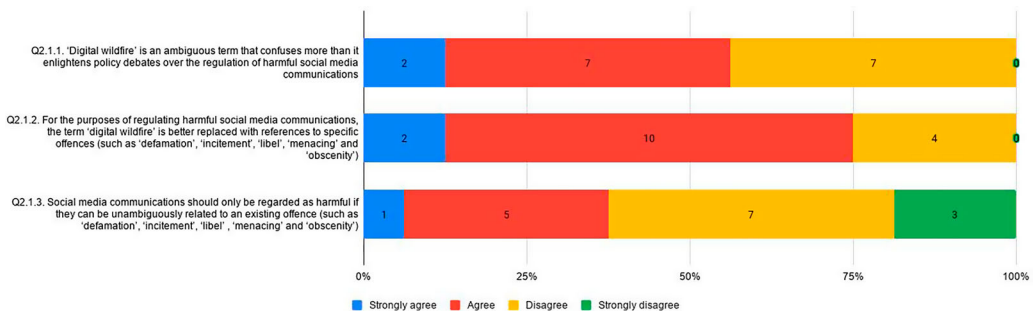

Figure 2. Clustering of opinion about the concept of 'Digital Wildfire'.

'online harms' whose regulation could, however inadvertently, result in the suppression of legitimate free speech. The other aspect concerns the adoption of more capacious definitions of online harm that can encompass injurious, albeit legal, communications but may, in turn, jeopardise liberal democratic freedoms, in so far as these definitions are used to suppress freedom of speech, for example, through 'no platforming'. The policy Delphi explored this strategic dilemma in greater depth in R2 and R3 by identifying the clustering of opinions about the political, as well as technical, feasibility of governing social media communications.

3.2. The technical and political feasibility of governing social media

In R1 participants were invited to provide free text responses to the questions: '*Insofar as Digital Wildfires ought to be controlled, how ought they to be controlled?*' and '*What, if any, limits would you place on freedom of speech in social media communications?*'. Their responses varied in preference for enforcement mechanisms and mirrored the spectrum of enforcement and non-enforcement considered in section 2.1., above:

Views on enforcement:

I am not sure if we need something else or just better enforcement of social media laws which sometimes is lacking. The police enforcement needs to be better. (U5) [prosecution][15]

Rather than place limits on freedom of speech, it can sometimes be better to allow these views to be publicly aired but ensure that they are appropriately and effectively challenged. This can then turn the tide of any digital wildfire, challenging hatred and incitement whilst not placing limits on freedom of speech. It is likely that if limits are placed on freedom of speech, those whose freedom of speech has been limited will only use this as further fuel for the wildfire. (I2) [disruption + self-regulation]

I think social media sites should have a responsibility to monitor increased network activity of posting a particular image/message etc (not sure if they have this capability yet) and they should be responsible for removing/blocking it in the first instance. (I1) [disruption + administrative penalties (for failure to monitor and thus disrupt)]

Regulation should facilitate strong media pluralism, so much so that the information from one media can be tested against information from another. (L3) [Licensing a broad media ecology]

Views on non-enforcement:

Digital wildfires, by nature, are not easily confined within national borders, and it would be difficult to limit online anonymity without compromising the usefulness of the Internet as a tool for whistle-blowers and political dissidents in repressive regimes. (U2) [Accommodation]

Social media users regulate by volume/strength of opinion which often lacks a factual or logical basis. Social rule by public opinion often does not uphold the values of a democratically elected government, can lead to 'mob rule' and vigilante-ism. (U1) [self-regulation by users]

Self-regulation and reporting from third party users will always benefit control of a situation and be able to recognise Digital Wildfires in their infancy. (I5) [self-regulation by users and platforms]

it is hard to regulate this type of behaviour through standard command-and-control regulation, nor is it even desirable to impose legal controls on speech that would be proportionate and necessary in a democratic society. Therefore, legal controls cannot work – leaving either technological controls or social/community-based controls. Technological controls are too simplistic or through design faults, either over filter or under filter. (L6) [disruption through technological controls, e.g. automated moderation, and/or user self-regulation]

The study drew on the R1 responses to prepare a questionnaire for R2 in which participants were presented with a series of enforcement and non-enforcement methods and asked to use a Likert scale to indicate how (1) technically feasible and (2) politically feasible they felt these to be. The clustering of opinion about this feasibility is illustrated in Figures 3 and 4.
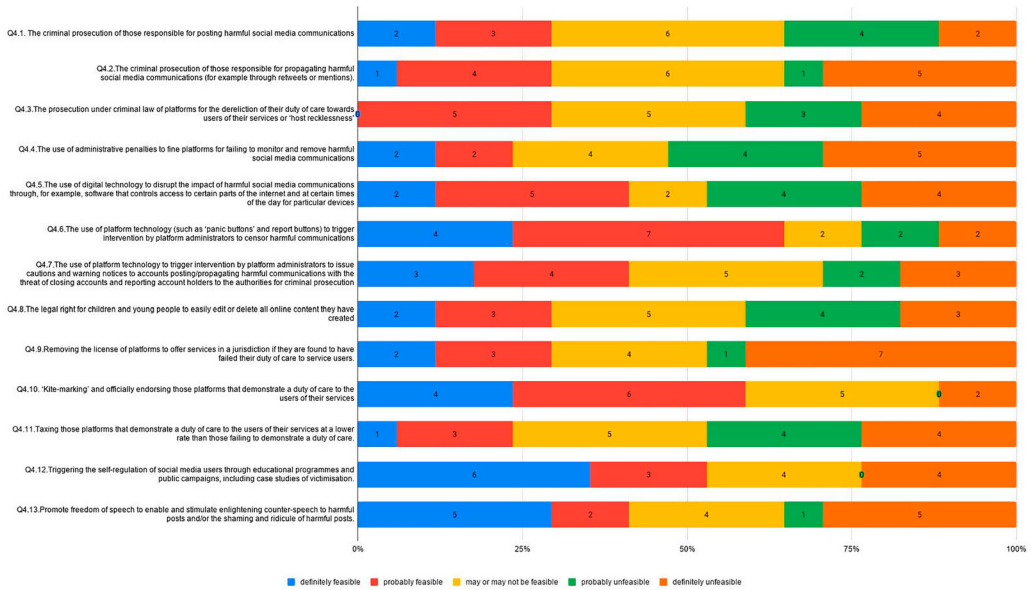
Figure 3. Clustering of opinion about the technical feasibility of regulating social media communications.

Only two of the strategies of regulation listed were regarded by a majority of the panel as both technically and politically feasible:

- Triggering self-regulation of social media users through educational programmes and public campaigns, including case studies of victimisation; and
- Use of platform technology (such as 'panic buttons' and report buttons) to trigger intervention by platform administrators to censor harmful communications.

Concerns were also expressed about the need to limit state control and avoid inhibiting freedom of speech. Respondents pointed out that, even if a measure is technically or politically feasible, it is not necessarily the 'right' thing to do and might not be effective in practice. Particular obstacles
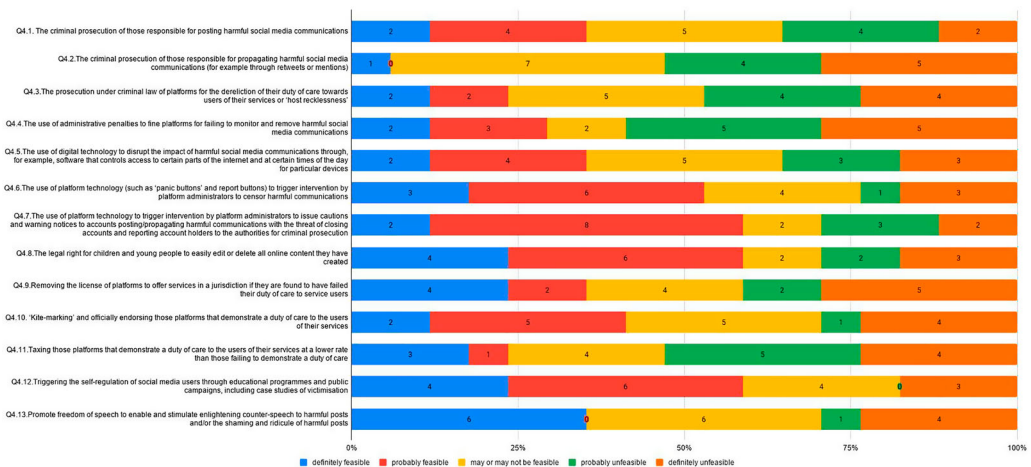

Figure 4. Clustering of opinion about the political feasibility of governing harmful social media.

referred to as limiting use of enforcement measures were: the ambiguous and subjective nature of harm as a concept; lack of an established understanding of what is meant by 'duty of care'; and absence of 'neutral' judges who might adjudicate instances of harm.

More specifically, it was argued:

It would not be possible to expect [social media platform] providers to stop wildfires as they developed given that they would have no means of taking a view on the truth or otherwise of a rumour. (U3)

… strong arm or overly excessive attempts to legislate [social media communications] could impact innovation of social platforms, and public backlash on over-intrusive government. (P2)

Licensing social media platforms is highly likely to lead to strong opposition from free speech groups and social media platforms. (L4)

I'm not convinced efforts to pressure platforms to conform to any one conception of 'care to the users' is ever appropriate. (L2)

Beneath these headline findings on the unfeasibility of enforcement, there were more subtle differences of opinion, which further refine understanding of strategic dilemmas between freedom of speech and harm reduction in the governance of social media:

Some high-level prosecutions of people who should know better (e.g. journalists, bloggers with large followings) may be acceptable if there is demonstrable harm. There would be less support for random prosecution of light-hearted gossip. (U3) [prosecution][16]

user-controlled restrictions on access to content are generally acceptable. For instance, parents can already use 'safe search' modes on Google. Search engines exist for children etc. (L4) [disruption]

This [Q.4.7. – platforms cautioning or suspending accounts flagged as 'harmful'] looks like the sort of approach which users might find acceptable as a means of controlling seriously damaging content - if that can be defined. (U3) [administrative penalties]

Uptake [Q.4.10. – of 'kite-marking' websites for high standards of customer care] will work for certain audiences, etc. In that narrow sense, this is very feasible. (L2) [licensing]

Ultimately self-regulation may not be seen as effective enough. (L5) [user self-regulation]

These outlier responses suggest prosecution may be effective in certain high-profile cases. They also suggest disruption can be both technically and politically feasible in relation to specific audiences (e.g. children) and when undertaken by specific authorities (e.g. parents) and even, possibly, by the state, although the political climate for this in liberal democracies may not yet exist, given concerns over unwarranted surveillance. Similarly, more bespoke forms of enforcement, such as suspension of social media accounts on proof of posting harmful content, as well as the use of kite-marking to reward conscientious platforms for moderating harmful communications, was regarded by some panellists as both technically and politically feasible. Finally, one panellist queried the adequacy for harm reduction of promoting user self-regulation through educational programmes and awareness-raising campaigns.

3.3. Forecasting: future scenarios for the governance of harmful social media

Given this clustering of opinion about feasibility of enforcement and adequacy of non-enforcement measures, R3 asked panellists to review the report of the R2 responses on feasibility and, in the light of these, return their views on the likelihood of five scenarios for the prospective regulation of harmful social media communications. The clustering of opinion about these forecasts is illustrated in Figure 5.

The overall findings from R3 corroborate panel-wide views expressed in R2. The weight of opinion across the panel suggests strategies of enforcement are unlikely to be the principal response in liberal democracies; instead greater emphasis and resources will be invested in enhancing user self-regulation or 'resilience'. However, promoting resilience is unlikely, in the judgement of the
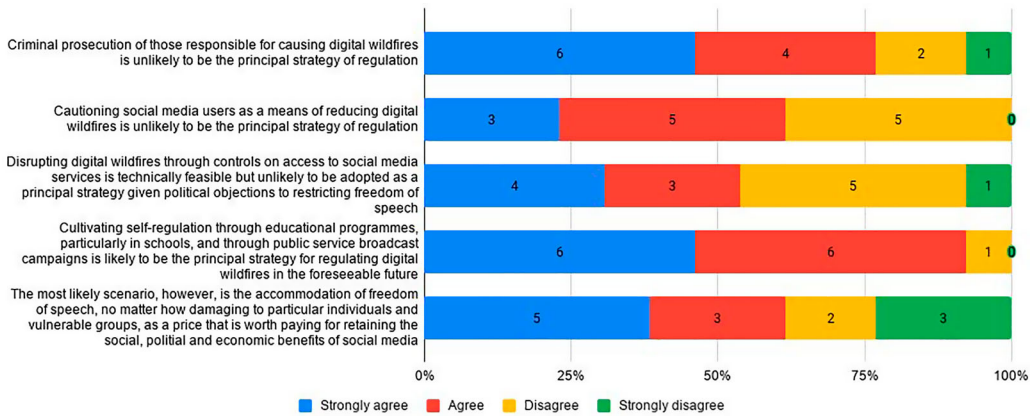
Figure 5. Future scenarios for governing harmful social media communications.

panel, to deliver on public demands for greater harm reduction, particularly for vulnerable groups, given the likely subordination of public protection from online harms to the perceived political and economic benefits of liberalised social media communication.

Analysis of free-text justifications provided by panellists, in particular of outlier responses, suggests a more nuanced anticipation of future scenarios for governance of harmful social media communications:

> While it is a good idea to prosecute individuals rather than holding platforms to account, anonymity will make this a difficult policy to pursue. (L7) [prosecution]

> … dismissing methods of controlling individual social media users as a way forward seems to be premature, particularly as there's much more that could be done with the technology itself. (L5) [cautioning]

> [disruption could be acceptable] if instigated by social media services themselves … given that private bodies can filter as much as they like and that they already engage in filtering e.g Facebook's news stream. (L3) [disruption]

> Whilst this [disruption through controls on access] could work in theory, these controls do not affect directly the person who started the wildfire; therefore they could start again as there is no actual consequence to the person. (I3) [disruption]

> [h]elping people overcome their own cognitive biases and errors in judgement is far better an approach than education and awareness campaigns. Defaults and architecture matter. (L6) [user self-regulation]

> Strongly agree with the sentiment [on the likelihood of accommodation of harms], but there are other solutions you have not yet contemplated. (L6) [accommodation]

These responses provoke further reflection on the effectiveness of other methods for reducing harmful social media communications without politically unfeasible constraints on freedom of speech. Here the argument turns on the recognition of social media platforms as sociotechnical systems that shape, and are shaped by, their users' behaviour.[17] The increasingly sophisticated set-up of social media platforms can shape users' behaviour – for instance, through default privacy settings, trusted contacts, panic/report buttons etc. Another aspect of the 'platform architecture' is the continuing advances made in automated moderation by censorship algorithms. Some argue that machine learning will become sufficiently nuanced, context-sensitive and culturally aware as to have the capacity to discriminate between abusive, satirical and sarcastic etc., communications, effectively censuring some whilst supporting the free articulation of others (Hendler and Berners-Lee 2010). Although these assumptions may be open to question (Edwards 2017, Kasparov 2017, Collins 2018), a key methodological point is that the policy Delphi format allows such

outlier viewpoints to be captured and presented for further discussion. In the policy Delphi, minority views are not knocked out between rounds, as there is no preoccupation with finding consensus. Instead, forecasting is structured around the identification of a breadth of scenarios.

4. Conclusion: social media and the policing of emergent technologies

Having presented the key findings of the Digital Wildfire policy Delphi, it is possible to adapt the study's theoretical propositions (sub-section 2.2., above) in light of insights gleaned from its three rounds of deliberation. These insights also highlight issues for the prospective research agenda in this rapidly evolving field of policing and regulation.

4.1. Enlightened self-regulation

Notwithstanding the questionable deterrent effect of prosecuting high-profile cases, scepticism over conventional policing responses to the spread of harmful content on social media is fuelling interest in prospects for self-regulation. Empirical studies of the propagation of content may have the capacity to identify what the role of users can be in shaping this propagation, including the curtailment of misinformation. Previous work has, for example, identified the capacity for 'counter speech' to curtail the spread of rumour in crisis scenarios (Procter *et al*. 2013) and hate speech in aggregated online content (Housley *et al*. 2018) and evidence to support this has subsequently been reported (Procter *et al*. 2019). As an understanding of self-regulation amongst high-profile online communities matures, counter speech is liable to play an increasingly significant role in the policing of harmful social media.

Panellists on the Digital Wildfire policy Delphi were less clear about how user self-regulation will evolve in the lower profile but higher volume abuses of social media found in mundane communications, especially amongst known vulnerable groups such as adolescents. School educational programmes on responsible social media use and other public awareness campaigns are already a major focus of policing in this field. However, epidemiological research suggests that limited opportunities for supervising the use of social media continues to fuel deteriorating trends in adolescent mental health (Kelly *et al*. 2018), questioning, in turn, the sufficiency of educational programmes and prospects for the sociotechnical disruption of online behaviour (see 4.3., below). As such, epidemiological studies of the consequences of social media use, the impact of resilience programmes in school and public awareness programmes about the propagation of abuse in mundane communications will continue to shape the research agenda in this rapidly evolving field.

4.2. Non-enforcement and 'digital gangsterism'

Findings suggesting that user self-regulation is limited amongst those social groups who are most vulnerable to online harms, imply a greater concern with the consequences of non-enforcement in governing social media. Hitherto, the principal focus has been on consequences of *laissez-faire* social media for the vulnerability of liberal democratic electorates and thus the integrity of the political process. For example, recent research identifies the potential of state actors to interfere in political events in other countries, as in the case of the Russian Internet Research Agency (Innes *et al*. 2019). By contrast, this study has focussed more on the mundane consequences of harmful social media communications in schools and workplaces. Whether in relation to mundane or exceptional instances of non-enforcement, issues of accommodation and collusion, regulatory capture, and the ownership and control of social media communications are necessarily part of the evolving research agenda on policing emergent technologies. These are themes that extend beyond the immediate purview of social media and democracy to other aspects of critical infrastructure. For example, insofar as commercial tech companies continue to enjoy the protection of 'trade secrecy' laws (Wexler 2018) for the software used to migrate the governance of healthcare,

energy and transport systems online, as in the 'smart cities' movement, this introduces another major vulnerability into public administration and a serious flaw in its democratic scrutiny and over-sight (Edwards and Calaresu 2018). In these terms, the policing of emergent technologies presents a major site of debate about potentially harmful, albeit legal, practices and a broadening out of the politics of criminalisation beyond the usual suspects of offline street crime to encompass the organ-isation of serious crimes online or, in the words of the House of Commons Select Committee inquiry into *Disinformation and Fake News*, the non-enforcement of 'digital gangsters' (DCMS 2019).

4.3. (Socio)technical fixes and hybrid human-machine learning

As conventional law enforcement is limited by the volume of social media communications and the capacity for human scrutiny and oversight, there is greater interest in the use of machine learning to police online harms. As noted by panellists on the Digital Wildfire policy Delphi, machine learning has the potential to produce false positives, especially given the global application of algorithms in diverse cultural contexts. In these terms, a new frontier in this research field is the prospects for hybrid human-machine learning, in which context-specific learning is advanced through the epi-sodic 're-training' of machines to detect and classify harmful communications. This, however, implies processes of human-machine learning that are explicit about the assumptions built into algorithms, including normative judgements of what constitutes 'harm'.

A further implication is that the use of machine learning needs to be open to democratic scrutiny and oversight and not covered by 'trade secrets' and intellectual property laws if their legitimacy and accountability is to be maintained (Wexler 2018). An equally thorny tension exists in the funding and implementation of machine learning where commercial tech companies with the requisite research and development capacity are enrolled into public administration. If trade secrecy is to be waived, this implies a significant reorientation of public expenditure around the policing of emergent tech-nologies undertaken with, or even entirely by, commercial tech companies.

4.4. Restrictions on freedom of speech

The facilitation by social media of asymmetries in electioneering presents liberal democracies with a central paradox: emergent technology, developed in part to revitalise civil society in these democra-cies, extending freedom of speech and removing editorial control in public discourse, now threatens the existential conditions of these democracies. The same mechanisms of freedom of speech without editorial control, enabled by social media, are driving the more mundane propagation of online harms, from trolling to revenge pornography to routine racist abuse. Mundane and extraordinary instances of digital wildfire need to be considered together in debates over the governance of social media, rather than methodologically bracketed-off from one another and treated as discrete topics of inquiry.

Considering mundane and spectacular cases of digital wildfires in tandem helps to further clarify the strategic dilemma for governance in this field of emergent technology. Considered in this broader institutional context, piecemeal exercises in user self-regulation and symbolic prosecutions of high-profile cases are unlikely to forestall escalation of online harms, including fake news, misin-formation and allied 'digital gangsterism'. Therefore, at what point will scandals over the agitation of social movements, manipulation of election campaigns and the epidemic of adolescent mental health problems propel liberal democratic authorities away from 'accommodation' and towards more draconian constraints on unmoderated social media communications?

4.5. Future research directions

Beyond the exhortation of social media users to police themselves or symbolic but ineffective prosecutions, a critical, albeit outlier, insight of the Digital Wildfire policy Delphi is that the poli-cing of online harms will take disruption as the fulcrum of liberal democratic responses to the

strategic dilemma of free speech and harm reduction. In this context, the frontier of research is likely to be the capacity and unintended effects of using increasingly nuanced hybrid human-machine learning to discriminate harmful from tolerable free speech, censoring the former and enabling the latter in sociotechnical systems that are open to democratic scrutiny and oversight.

Notes

1. See World Economic Forum (2013).
2. 'Big brands fund terror through online adverts: Household names unwittingly pay extremists and pornographers.' *London Times*, 9th February 2017: https://www.thetimes.co.uk/article/big-brands-fund-terror-knnxfgb98 [Accessed 6 October 2017].
3. 'Inside Russia's Social Media War on America.' *Time Magazine*, 18th May 2017. Available from: http://time.com/4783932/inside-russia-social-media-war-america/ [Accessed 6 October 2017].
4. In which a tweet by Sally Bercow, wife of the Speaker of the House of Commons, libelled the Conservative Peer Lord McAlpine, suggesting he was a paedophile, at: https://www.theguardian.com/politics/2013/may/24/sally-bercow-tweet-libelled-lord-mcalpine.
5. John Raymond Nimmo and Isabella Kate Sorley received prison sentences, under the Communications Act 2003, for sending menacing communications about the feminist campaigner Caroline Criado-Perez, https://www.judiciary.uk/wp-content/uploads/JCO/Documents/Judgments/r-v-nimmo-and-sorley.pdf.
6. https://www.theguardian.com/society/2019/jul/26/racist-facebook-troll-jailed-for-abuse-of-female-politicians-gerard-traynor.
7. A problem unlikely to be allayed by further legal reforms that extend the criminalisation of harmful online communications, as in the Law Commission's proposed reform of malicious communications laws in England and Wales to encompass new offences such as 'group harassment' and 'cyber-flashing' (Law Commission, 2020).
8. https://www.news-medical.net/news/20200318/Social-media-giants-join-forces-to-tackle-spread-of-fake-coronavirus-news.aspx.
9. The concept of 'produsers' acknowledges the active role of users of social media in simultaneously producing as well as consuming public communications. The allied concept of 'read/write' technology captures this distinctive quality of social media communications, as enabled by the interactional functionality of Web2.0, distinguishing them from the more passive consumption of other broadcast and print media. It is precisely this read/write facility for producing communications globally and in real-time that has created a new condition for the governance and regulation of communications that are outflanking national policing and criminal justice capacities.
10. See address by Baroness Kidron to the Digital Wildfires project dissemination event, Digital Catapult, London 12th January 2016, see: https://www.youtube.com/watch?v=VDX34XBgRLA.
11. 'Many, on both sides [of the left and right in American politics] believe that government should actively regulate the moderation of social media platforms to attain fairness, balance, or other values. Yet American law and culture strongly circumscribe government power to regulate speech on the internet and elsewhere' (Samples, 2019, p. 1). Whilst social media companies themselves may restrict individual speech for transgressions against 'community rules' for using the platform in question, The First Amendment of the US Constitution provides powerful grounds for redress against such restrictions on free speech and the US Congress has also protected tech companies from 'intermediary liability for speech that appears on their platforms' (Samples, 2019, p. 1).
12. Examples given of digital wildfires included: allegations connecting high profile individuals to historic child sexual abuse crimes; Gamergate; and tweets about Martin Shkreli (Turing Pharma); rumours about the death of Whitney Houston; Justine Sacco's 'racist' tweet about HIV and Africa; comments on Reddit about the identity of the Boston bomber; the online 'shaming' of individuals.
13. 'L6' = Lawyer number 6 on the panel, 'L4' = Lawyer number 4, 'U4' = User/scientist number 4, 'P2' = Social Media Platform representative number 2, 'I2' = Institutional representative number 2 on the panel.
14. 'Q2.1.1' etc., refers to the original question identifier. In this instance, questionnaire for the second round, section one, question one.
15. [prosecution], [disruption] etc., are the axial codes used to interpret R1 responses in terms of the spectrum of enforcement and non-enforcement.
16. Again, these outlier views can be coded in terms of the spectrum of enforcement and non-enforcement measures used to conceptualise the problem of governing social media communications (see sub-sections 2.1. and 2.2., above).
17. 'Put simply, the sociotechnical system perspective contends that organisations are made up of people that produce products or services using some technology, and that each affects the operation and appropriateness of the technology as well as the actions of the people who operate it' (Pasmore et al, 1982, p. 1182).

Acknowledgements

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

ORCID

Adam Edwards 🆔 http://orcid.org/0000-0002-1332-5934
Helena Webb 🆔 http://orcid.org/0000-0002-4303-7773
William Housley 🆔 http://orcid.org/0000-0003-1568-9093
Roser Beneito-Montagut 🆔 http://orcid.org/0000-0001-5967-4307
Rob Procter 🆔 http://orcid.org/0000-0001-8059-5224
Marina Jirotka 🆔 http://orcid.org/0000-0002-6088-3955

References

Ahmed, W., *et al.*, 2020. COVID-19 and the 5G conspiracy theory: social network analysis of Twitter data. *Journal of Medical Internet Research*, 22 (5), e19458. doi:10.2196/19458.
Beran, T. and Li, Q., 2007. The relationship between cyberbullying and school bullying. *Journal of Student Wellbeing*, 1 (2), 15–33.
Collins, H.M., 2018. *Artifictional Intelligence: Against Humanity's Surrender to Computers*. Cambridge: Polity Press.
Crown Prosecution Service. 2018. *Social media – guidelines on prosecuting cases involving communications sent via social media*. Available from: https://www.cps.gov.uk/legal-guidance/social-media-guidelines-prosecuting-cases-involving-communications-sent-social-media [Accessed 4 July 2020].
DCMS. 2019. Disinformation and 'fake news': final report of the house of commons digital, culture, media and sport committee. Available from: https://www.parliament.uk/business/committees/committees-a-z/commons-select/digital-culture-media-and-sport-committee/news/fake-news-report-published-17-19/.
de Loe, R.C., *et al.*, 2016. Advancing the state of policy Delphi practice: a systematic review evaluating methodological evolution, innovation, and opportunities. *Technological Forecasting and Social Change*, 104, 78–88.
Depoux, A., *et al.*, 2020. The pandemic of social media panic travels faster than COVID-19 outbreak. *Journal of Travel Medicine*, 27 (3), taaa031. doi:10.1093/jtm/taaa031.
Edwards, A., *et al.*, 2013. Digital social research, social media and the sociological imagination: surrogacy, augmentation and re-orientation. *International Journal of Social Research Methodology*, 16 (3), 245–260.
Edwards, A., 2017. Big data, predictive machines and security: the minority report. In: M. R. McGuire and T. J. Holt, ed. *The Routledge Handbook of Technology, Crime and Justice*. Abingdon: Routledge, 451–461.
Edwards, A. and Calaresu, M., 2018. Smart cities and security. *City, Territory and Architecture [online]*, 5, 19. doi:10.1186/s40410-018-0089-1.
Edwards, A. and Gill, P., 2002. Crime as enterprise? The case of transnational organised crime. *Crime, Law and Social Change*, 37 (3), 203–223.
Gill, P., 2002. Policing and regulation: what is the difference? *Social and legal studies*, 11 (4), 523–546.
Hendler, J. and Berners-Lee, T., 2010. From the semantic web to social machines: a research challenge for AI on the World Wide Web. *Artificial Intelligence*, 174, 156–161.
Housley, W., *et al.*, 2014. Big and broad social data and the sociological imagination: a collaborative response. *Big Data & Society [online]*, 1, 2. doi:10.1177/2053951714545135.
Housley, W., *et al.*, 2017. Membership categorisation and antagonistic Twitter formulations. *Discourse & Communication*, 11 (6), 567–590.
Housley, W., *et al.*, 2018. Interaction and transformation on social media: the case of Twitter campaigns. *Social Media and Society*, 4 (1), 1–12.
Innes, M., Dobreva, D., and Innes, H., 2019. Disinformation and digital influencing after terrorism: spoofing, truthing and social proofing. *Contemporary Social Science [online]*. doi:10.1080/21582041.2019.1569714.

Jarvis, J. 2018. Platforms are not publishers: the essential value of the internet is conversation, not content – and journalists need to embrace it. *The Atlantic* [online]. Available from: https://www.theatlantic.com/ideas/archive/2018/08/the-messy-democratizing-beauty-of-the-internet/567194/.

Juvonen, J. and Gross, E.F., 2008. Extending the school grounds? Bullying experiences in cyberspace. *Journal of School Health*, 78, 496–505.

Kasparov, G., 2017. *Deep thinking: where machine intelligence ends and human creativity begins*. London: John Murray.

Kelly, Y., *et al*., 2018. Social media use and adolescent mental health: findings from the UK millennium cohort study. *EClinical Medicine*, 6, 59–68.

Law Commission, 2020. *Harmful online communications: the offences. consultation paper 248*. London: Law Commission.

Layder, D., 1998. *Sociological practice: linking theory and social research*. London: Sage.

Lowe, T. and Innes, M., 2012. Can we speak in confidence? Community intelligence and neighbourhood policing v2.0. *Policing and Society*, 22 (3), 295–316.

McNair, B., 2018. *An introduction to political communication*. London: Routledge.

Parliament House of Lords, 2019. *Regulating in a digital world. HL 2017-19 (299)*. London: The Stationary Office.

Pasmore, W., *et al*., 1982. Sociotechnical systems: A North American reflection on empirical studies of the seventies. *Human Relations*, 35 (12), 1179–1204.

Procter, R., *et al*. 2019. A study of cyber hate on Twitter with implications for social media governance strategies. *In*: Proceedings of the Workshop on Trust and Trust Online, 4–5 October, London.

Procter, R., 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), 197–214.

Samples, J., 2019. Why the government should not regulate content moderation of social media. *Policy Analysis, Cato Institute*, 865, https://www.cato.org/sites/cato.org/files/pubs/pdf/pa_865.pdf.

Sumsion, T., 1998. The Delphi technique: an adaptive research tool. *British Journal of Occupational Therapy*, 61 (4), 153–156.

Turoff, M., 1970. The design of a policy Delphi. *Technological Forecasting and Social Change*, 2, 149–171.

Webb, H., *et al*., 2015. Digital wildfires: hyper-connectivity, havoc, and a global ethos to govern social media. *Computers and Society*, 45 (3), 193–201.

Webb, H., *et al*., 2016. Digital wildfires: propagation, verification, regulation, and responsible innovation. *ACM Transactions on Information Systems (TOIS)*, 34 (3), 1–23.

Wexler, R., 2018. Life, liberty, and trade secrets: intellectual property in the criminal justice system. *Stanford Law Review [online]*, 1343, doi:10.2139/ssrn.2920883.

World Economic Forum (2013) Digital Wildfires in a Hyperconnected World. World Economic Forum Global Risks Report 2013. Available from: http://reports.weforum.org/global-risks-2013/risk-case-1/digital-wildfires-in-a-hyperconnected-world/

Zubiaga, A., *et al*., 2018. Detection and resolution of rumours in social media: a survey. *ACM Computing Surveys [online]*, 51, 2. doi:10.1145/31616.