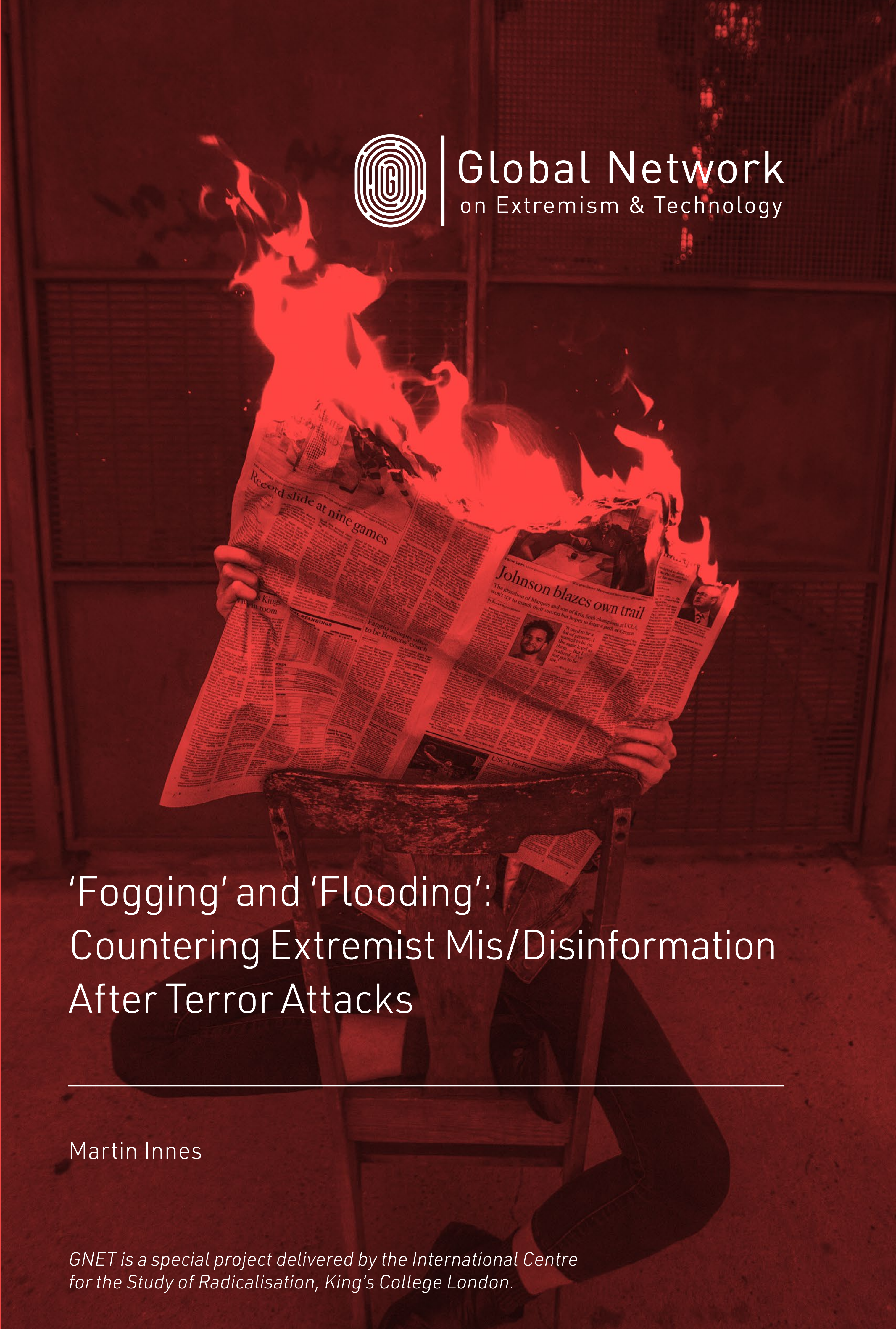

Global Network
on Extremism & Technology

‘Fogging’ and ‘Flooding’: Countering Extremist Mis/Disinformation After Terror Attacks

Martin Innes

GNET is a special project delivered by the International Centre for the Study of Radicalisation, King’s College London.

The author of this report is Martin Innes, Director of the Crime and Security Research Institute, and Universities' Police Science Institute at Cardiff University

The Global Network on Extremism and Technology (GNET) is an academic research initiative backed by the Global Internet Forum to Counter Terrorism (GIFCT), an independent but industry-funded initiative for better understanding, and counteracting, terrorist use of technology. GNET is convened and led by the International Centre for the Study of Radicalisation (ICSR), an academic research centre based within the Department of War Studies at King's College London. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing those, either expressed or implied, of GIFCT, GNET or ICSR.

CONTACT DETAILS

For questions, queries and additional copies of this report, please contact:

ICSR
King's College London
Strand
London WC2R 2LS
United Kingdom

T. **+44 20 7848 2098**
E. **mail@gnet-research.org**

Twitter: **@GNET_research**

Like all other GNET publications, this report can be downloaded free of charge from the GNET website at www.gnet-research.org.

Executive Summary

Social media and associated changes to the media ecosystem have profoundly impacted upon the dynamics of public sense-making and understanding in the aftermath of terror attacks. This analysis looks at how and why misinformation and disinformation arises in such situations, what impacts it has and what might be done to manage and mitigate its effects.

The main focus of the report is on introducing three innovative concepts intended to help us to interpret and understand these processes of social reaction:

- ‘Fogging’ is an effect that arises from constructing and communicating multiple explanations and interpretations of the events in question. These accounts can be more or less plausible. The purpose of transmitting these alternative versions of reality is not necessarily that they should be widely believed; they simply have to be sufficient to induce a sense of doubt and complexity about the underlying causes. The effect is to create a miasma of competing and contrasting accounts and explanations in the information space, such that public audiences do not quite know what to believe happened or why, or which sources can be trusted.

- ‘Flooding’ is a related but distinct informational effect that involves dominating an information space with a particular misinforming or disinforming message. This involves reposting the message in high volumes and frequently across platforms to make it highly visible and likely to be encountered repeatedly by audience members engaging with the event or issue of concern. Under a general condition of fogging, then, ‘flooding the zone’ with a particular distorting or deceptive message constitutes a specific influencing effect that reinforces and reproduces the wider condition of which it is a part.

- ‘Surfacing’ refers to some specific techniques of persuasion used to establish a patina of plausibility to the alternative narratives being constructed and that are used to fog and flood the zone of influence. Some key examples of ‘surfacing’ mis/disinforming messages are: by claiming to be an eyewitness; by using visual images that are claimed to be of the scene but are not really; by drawing attention to other sources online proffering alternative accounts.

Framing these concepts and their application is a recognition that, although in the academic literature a ‘bright-line’ distinction is routinely drawn between misinformation and disinformation, their empirical manifestations often tend to be rather more complex and contingent. Misinformation is typically defined as the unintentional transmission of misleading information. Contrastingly, disinformation is held to involve an intentional act to distort or deceive. However, in situations such as terror attacks that are rapidly evolving and marked by high degrees of uncertainty and imperfect information, it is often hard to divine the intent behind a particular message or communication. Moreover, it is

commonplace for a deliberately misleading message to be amplified and redistributed unwittingly by actors who sincerely believed that it was accurate, and vice versa. Thus raising difficult questions about whether such episodes should properly be defined as mis- or disinformation.

Reflecting such contingencies and ambiguities, herein we use the concept of ‘mis/disinformation’ to articulate how the main analytic focus is upon distilling some aggregated informational effects, that frequently arise out of a blend of false messages, some of which were deliberately manipulative, where others had more benign origins. This is reasonable given how each of the three main concepts developed is concerned with capturing an aggregated meso-level consequence, rather than the specific effects triggered by an individual message.

To develop these concepts and demonstrate how they illuminate the dynamics of public reactions to terror attacks, some empirical data is introduced drawn from a wider programme of research using social media data to intensively study the aftermath of high-profile terror attacks. Some of the key episodes reported on include:

- How following the bombing of the Manchester Arena, images posted on Twitter from inside the venue were challenged and disputed, with allegations they were a part of a ‘false flag psyop’. In turn, these accounts created a conducive environment for other mis/disinformation narratives that had serious consequences, including one claiming that there was another attacker at Oldham hospital.

- The misidentification of Abu Izzadeen as the perpetrator of the Westminster Bridge attack in 2017 and how this was amplified by far-right groups to promote their ideological narratives. Far-right groups continued to use this misidentification even after it had been debunked.

One particular influence vector identified by the analysis relates to how mis/disinformation constructed and transmitted by fringe websites claiming journalistic credentials and their social media presences, can have an impact upon the behaviour of more mainstream media outlets, as they compete for ‘scoops’ and being first to report on the details, even when they know the integrity of the substantive material might be compromised.

More broadly, this report makes the case for the importance of those studying different kinds of online harms to trade and exchange ideas so that researchers can build up a ‘richer picture’ and develop a more nuanced understanding of the intersections and interactions between new technologies and social harms.

Contents

1 Introduction

This report explores how and why mis/disinformation develops in the wake of terror attacks and the ways it is used by extremist groups to attempt to shape public understanding and political responses. These uses include extremist sympathisers engaging in information manipulation and obfuscation as part of their attempts to explain or justify the violence, as well as distorting and deceptive messaging designed to marginalize or stigmatize other social groups. Having presented evidence and insight about the construction of these messages, the discussion also looks at the policy and practice options in terms of ‘what works’ with regard to managing and mitigating any such messaging and the harms it seeks to induce.

It is now largely taken for granted that social media and the wider changes to the media ecosystem with which it is associated have had profoundly disruptive and transformative impacts upon the institutional and interactional ordering of society. But while very few social and political commentators would contest the general tenor of this assertion, it is increasingly clear that the effects of social media upon patterns of communication and knowledge are complex, especially with respect to specific policy and practice domains.[1] One such domain is political violence and the countermeasures intended to limit its effects.

Terrorist violence is fundamentally a form of communicative action. The violent act is designed to deliver a message in pursuit of a political objective or in response to some grievance. The proliferation and diversification of social media has altered the processes of social communication associated with terror attacks.[2] First, it has changed the conduct of the violent act itself with increasing numbers of assailants designing the delivery of attacks in ways that encourage social media dissemination, for example in terms of promoting their manifestos and/or livestreaming the attacks themselves. Second, it has impacted the processes of social reaction to terror attacks, as supporters and ideological opponents of the perpetrator engage in ‘framing contests’ to try to establish a public definition of the situation in terms of how the violence is interpreted and understood. A third strand of influence relates to the social control responses of governments, police and intelligence agencies, who increasingly have to think about the effects of social media messaging on their control strategies, in terms of not only the practicalities of any investigation but also public reassurance.

Finally, social media has afforded new ways of studying patterns of social reaction in the aftermath of terror attacks (this is especially significant from the perspective of this paper). In particular, the streaming quality of many social media feeds has opened up new ways of capturing what happens in the minutes, hours, days and

[1] Margetts, H., John, P., Hale, S. & Yassera T. (2016) *Political Turbulence: How Social Media Shape Collective Action*, Princeton: Princeton University Press.

[2] Innes, M. (2020) “Techniques of disinformation: Constructing and communicating ‘soft facts’ after terrorism”, *British Journal of Sociology*, vol.71 no.2: pp.284–99.

weeks following an event.[3] This is significant in that, until fairly recently, relatively little research attention had been directed towards analysing processes of social reaction in the aftermath of terror attacks. Compared with the amount of attention that has been directed to the upstream issue of violent extremist radicalisation, little work has focused upon the immediate post-attack situation and how this shapes public perceptions and understanding. In part, this reflects the challenges of tracking and tracing the dynamics of public opinion, especially in terms of thinking about how distinct audience segments may display substantively different response patterns. However, social media provides a source that is simultaneously rich in detail, but also available at scale.

Reflecting this trajectory of development, over the past five years or so there has been a growing literature using social media data to illuminate different facets of what occurs in the post-violence moment. For example, Randall Collins initially used his theoretical work on the time dynamics of conflict to argue that there will be moments of ‘collective emotion’ and thus heightened risks of further hate crime and violence in the period following a major attack.[4] Using empirical data collected following the murder of Lee Rigby by Islamic fundamentalists in London in 2013, Roberts et al. (2018) found key elements of Collins’ theory to be supported.[5] Developing this insight that there may be ‘reaction patterns’ to the organisation of public responses to terrorism, further work suggested that there were ‘ten Rs’ of reaction. This included, of particular salience to this report, what was labelled ‘rumouring’, the transmission of speculative and unverified information of uncertain origin.[6]

Cast in this light, social media functions as what the sociologist Donald Mackenzie (2008) dubbed both ‘an engine’ and a ‘camera’.[7] This is because it drives changes in the causes, conduct and consequences of terrorism while simultaneously it functions as a source of data available to researchers to capture the details and intricacies of what happens in terms of the evolution of public perceptions and sentiments, and how these vary across particular audience groups and segments.

The focus of this report is upon one specific facet of this public interpretation and sense-making process, whereby groups author and/or amplify mis/disinformation to obfuscate and obscure particular definitions of the situation. Often this occurs as they try to subvert and contest interpretations that they do not like, that run counter to their group’s ideological values and objectives. The mis/disinforming messages can originate ‘organically’ out of the chaos and confusion that arises immediately post-attack or be more deliberately and purposively ‘manufactured’. They are authored and amplified to encourage and motivate supporters and sympathisers, while simultaneously triggering reactions from ideological adversaries.

[3] Innes M. et al. (2018) “From Minutes to Months: A Rapid Evidence Assessment of the Impact of Media & Social Media During & After Terror Events”, Public Safety Canada.

[4] Collins, R. (2012) “C-Escalation and D-Escalation: A Theory of the Time-Dynamics of Conflict”, *American Sociological Review*, vol.77: pp.1–20.

[5] Roberts, C., Innes, M., Preece, A. and Rogers, D. (2018) “After Woolwich: Analysing open source communications to understand the interactive and multi-polar dynamics of the arc of conflict”, *British Journal of Criminology*, vol.58 no.2: pp.434–544

[6] Innes, M., Roberts, C., Preece, A. and Rogers, D. (2018) “Ten Rs of social reaction: Using social media to measure the post-event impacts of the murder of Lee Rigby”, *Terrorism and Political Violence*, vol.3 no.3: pp.454–74.

[7] Mackenzie, D. (2008) *An Engine Not a Camera: How Financial Models Shape Markets*, Cambridge, Massachusetts: MIT Press.

Disinformation is where communicating false or misleading information is intentional. Misinformation, by contrast, involves the unwitting transmission of such material. Because divining intent for such communicative actions with any degree of analytic confidence is increasingly difficult, especially in post attack situations that are so uncertain and riven with imperfect information, throughout this discussion we will simply refer to mis/disinformation. As a concept this formulation rather neatly captures some of the ambiguities and contingencies that pertain to the messaging that various actors transmit in their attempts to influence how others perceive and understand the causes and consequences of the violence that has occurred. It also allows the discussion to avoid getting hung up on defining whether intent is present or not, in order that the analysis can attend more to the impacts and consequences that flow from the stream of distortions and deceptions identified.

Framed in this way, the two core techniques centred by the analysis can be defined as follows:

- 'Fogging' involves constructing and communicating multiple explanations and interpretations of the events in question. These can be more or less plausible. The purpose of transmitting these alternative versions of reality is not necessarily that they should be widely believed, but simply sufficient to induce a sense of doubt and complexity about the underlying causes. This creates a miasma of competing and contrasting accounts and explanations in the information space, such that public audiences don't quite know what to believe happened or why.

- 'Flooding' involves ensuring that the information space is dominated by a particular mis/disinforming message. Reposting a message in high volumes and frequently across platforms makes it highly visible and likely to be encountered repeatedly by audience members engaging with the event or issue of concern. Under a general condition of fogging, then, 'flooding the zone' with a particular distorting or deceptive message constitutes a specific influencing tactic that reinforces and reproduces the wider condition of which it is a part.

In the process of elaborating how fogging and flooding work as part of attempts to influence public perceptions and understanding, a third concept of 'surfacing' is also introduced:

- 'Surfacing' refers to some specific techniques of persuasion used to establish a patina of plausibility to the alternative narratives being constructed and used to fog and flood the zone of influence.

These concepts are developed through analysis of empirical data originally collected as part of a long-term international research programme focused on understanding public reactions to terror attacks.

The central innovation and contribution to knowledge of developing these three constructs is to connect recent developments in the study of social reactions to terrorism with the rapidly growing literature on the construction and communication of mis/disinformation.[8] It is an approach that helps to illuminate some of the complexity of public interpretation and sense-making with regard to fabricating the political and societal meanings and significance that are ascribed to particular incidents. Ultimately, it provides a deeper, more nuanced understanding of terrorist and extremist propaganda, and the communicative actions that the authors of such material perform as they seek to garner social support for their cause while simultaneously degrading and discrediting opposing ideas and values.

The next section provides a brief overview of the procedures via which the empirical materials underpinning this analysis were collected. Extracts from several different case studies are then introduced to map out the contours of the first key concept, fogging. As part of this, the specific role of surfacing techniques is described, before a similar empirically guided account of flooding the zone is provided. This is then followed by an analysis of some strategic and tactical options for disrupting and degrading the impacts of such techniques of disinformation, including the different stakeholders who might potentially leverage them.

[8] See, for example, Benkler, Y., Faris, R. and Roberts, H. (2018) *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics*, New York: Oxford University Press; Woolley, S. and Howard, P. (2019) *Computational Propaganda: Political Parties, Politicians and Political Manipulation on Social Media*, Oxford: Oxford University Press.

2 Research Design and Data

As outlined above, this paper’s main focus is on developing the two key concepts. To inform this, empirical materials from a wide-ranging research programme are drawn upon for illustrative purposes. This broader programme involves cross-platform social media data collected in the wake of multiple terrorist incidents, including: the four terror attacks that occurred in the UK in 2017; the series of attacks that took place in France and Germany between 2015 and 2020; the murder of Jo Cox MP in 2016; and the murder of Fusilier Lee Rigby in 2013. For all of these, data tracking the evolution and adaptations in the processes of public reaction was systematically collected from multiple platforms and stored, including content showing how different extremist groups were competing to establish definitions of the situation aligned with their ideological values and to amplify the sense of harm induced.

The empirical materials are intended to support the conceptual development work, rather than be the principal focus of the analysis. Thus, it is sufficient to provide an overview of how they were collected, as opposed to a fully detailed, in-depth methodological account.

Data was collected using a suite of instruments and tools reflecting how different packages provide particular affordances; no single tool can do it all when it comes to social media monitoring and analysis. The key tools were: Sentinel, Cardiff University’s bespoke package, which comprises a number of innovative data collection and processing apps; GDELT and Webhose global media aggregators, which collect material from international mass media outputs associated with the incidents of interest; Brandwatch, which provides datasets from the Twitter API; CrowdTangle, for public Facebook and Instagram data; and TGStat, for Telegram metrics. The data was supplemented by manual scraping of sources such as VK, Gab, TikTok and Parler where appropriate. Importantly, all of the data was publicly available and open source.

As intimated above, a lot of the relevant data was collected in reaction to the occurrence of terrorist and extremist events of interest. Many of the packages provide for an ongoing monitoring capability based around a series of user-defined keywords that are used to scan social data streams for (for example: ‘bomb’; ‘explosion’; ‘terrorist’). When an attack occurred, the researchers would rapidly refine these keywords to focus on collecting data of interest until public engagement had subsided. The intention was to build up a data series that would enable a comparative case study research design, enabling studies of the common patterns that exist in terms of social reactions to terror attacks or at least how they present on social media.

As a methodology, there are some limitations that are worth drawing out. First, tuning the collection in the early stages of an event does introduce some variation over time in terms of what is and is not collected. That said, for many of the incidents where data was collected, we did detect the first public social media messages transmitted as well as many of the other important ones. There is, though, a ‘memory hole’ problem that results from the fact that the major social media companies have become increasingly interventionist in removing harmful content from their platforms, which they now attempt to do rapidly and at scale in the aftermath of almost every terrorist incident. When they remove accounts for breaching their terms of service or community standards for harmful content, all the messages are removed as well. The Sentinel platform allows us to offset this to some extent, for Twitter and Reddit in particular, but it is an issue in terms of the integrity of the overall datasets and the insights that can be derived. Finally, as mentioned previously, as part of an approach to ethical and transparent research, we only use publicly available, open source data and do not collect or analyse data from platforms such as WhatsApp or any groups or forums that one has to request to join.

3 Fogging

In recent years there has been rapid growth in the use of social media to track and trace patterns of public reaction in the aftermath of terror attacks. One of the key findings of this work is that the period immediately following an attack is especially vulnerable to the transmission of misinformation and disinformation.

Although the prevalence of mis/disinformation is often especially acute in the immediate aftermath, it can be a more persistent issue as well, extending over a considerable period of time. Herein we are especially interested in the consequences of the circulation of misleading communications and introduce the concept of fogging to capture this. It is a notion intended to articulate how the multiple competing accounts, interpretations and explanations transmitted via a variety of sources and channels aggregate to a ‘miasmatic quality’: while they may blur and obscure the reality of what has happened and make it more difficult to define with clarity, they do not obliterate it entirely.

‘Natural’ Fogging

It is well documented that the period immediately following a terror attack is defined by acute uncertainty about what has happened and how it should be interpreted and understood. As a consequence, rumours and misinformation are frequently transmitted as people individually and collectively struggle to make sense of what is unfolding and to establish a coherent definition of the situation. This we can label ‘natural fogging’ to convey how it is an almost inevitable and unintended outcome of the struggle to understand a shocking and traumatic occurrence. As will be discussed in the next section, it can be contrasted with the idea of a more purposeful ‘engineered fog’.

An illustration of the communicative dynamics of natural fogging is provided by some of the messaging in the immediate aftermath of the 2017 bombing of the Manchester Arena. In different ways these messages served to confuse and complicate the public understanding of what was happening. There were multiple instances of mis/disinformation transmitted, several of which are discussed by Innes et al. (2020). To take just one example: shortly after the attack on 22 May 2017, photographs started to be shared on social media claiming to show the doors at the Metrolink entrance into the main arena hall in the immediate aftermath of the bombing.[9] At around the same time, other messages were being transmitted on Twitter that the loud noise was actually just some balloons popping or a speaker exploding, which is relevant to how the photographs were interpreted by some. It is possible that the tweets were posted by people nowhere near the venue; they clearly constituted either misinformation or disinformation.

[9] This entrance connects the arena venue with the Manchester Victoria railway station.

One image in particular, taken a short distance outside the doors looking in, showed bodies lying on the floor. This photo, which was taken by a journalist, appeared on social media only two minutes after Greater Manchester Police tweeted that it was responding to an emergency incident.[10] In another photo, emergency services personnel were shown alongside pixelated images of bodies.

The authenticity of these photos was immediately contested on social media. Within an hour of the start of the incident a variety of messages were circulating that the images were from a training exercise, although the type varied (police or army) as did the date (alternatively ‘last year’, ‘recent’ and ‘two years ago’). No one provided definitive details of the training event, nor was there any official response to these claims. However, some tweets referred to the police terrorism exercise held at the Manchester Trafford Centre on 10 May 2016, which had attracted media controversy.[11]

From the point of view of this paper, we are interested not just in the emergence of these unwarranted assertions and the ways they induced a fogging effect, but also how they were given a veneer of plausibility, through what can be labelled ‘surfacing techniques’. Looking across the accounts that claimed the images were fake, it can be observed that they invoke a series of individual and collective surfacing techniques. Specifically:

- Some of the messages did not contest the visual evidence of the photos, but rather how they were being interpreted, suggesting they were taken at a different time and/or place.

- Others focused upon particular details within the images and extrapolated from these. For example, that the blood patterns did not look genuine.

- Some offered motivations as to why the original posters of the images should not be viewed as reliable. They were: ‘attention seeking to boost their online follower numbers’, engaged in a ‘false flag psyop’ or taking advantage of photo-editing software to manipulate the scene.

- Finally, there was a ‘social proofing effect’ that derived from the fact that these messages from multiple accounts were retweeted and reposted in relatively high volumes over a short time. This meant that those tracking the unfolding event online were quite likely to encounter the mis/disinformation via several different sources, rendering it more believable. The effect of this can be evidenced by comments online expressing sentiments such as “there are too many people online saying this is fake for it to be true.”

At first sight, this particular flurry of claims did not in and of itself appear to have much of a material impact upon public sense-making and understanding. However, this is to dismiss the more subtle influence it had on these collective sense-making processes. Because it happened so soon after the incident, it had a ‘priming and framing’

[10] “Manchester Bombing: What We Don’t Know,” *TruePublica*, 24 May 2017, https://truepublica.org.uk/united-kingdom/manchester-bombing-what-we-dont-know.

[11] Frances Perraudin, “Police Apologise For ‘Allahu Akbar’ Use in Mock Manchester Attack,” *The Guardian*, 10 May 2016, https://www.theguardian.com/uk-news/2016/may/10/police-apologise-for-allahu-akbar-use-in-mock-manchester-attack.

effect, inasmuch as it visibly illuminated how there was false and misleading information in the communications system, such that doubting the provenance of online messaging appeared a wise strategy. It induced a sense of doubt and uncertainty about who and what to believe.

The consequences of this transpired a short time later when a message appeared on Facebook, subsequently reposted rapidly and in high volumes on Twitter, that claimed:

> *DO NOT COME OLDHAM HOSPITAL IM CURRENTLY LOCKED INSIDE… MAN OUTSIDE WITH GUN*

The use of capitals imbued the original Facebook post with a clear sense of both urgency and emergency. We do not have an exact time for the post, but detected it as a result of a screenshot the messenger sent to a friend at 00:22. It is not clear whether the origins of this claim should properly define it as misinformation or disinformation, but either way, it had serious material consequences. The prospect that there could be a second active terrorist in the area seemingly contributed to the decision to hold the majority of ambulance and fire crews at the outer security cordon, rather than allowing them in to the scene to treat victims.

For the purposes of clarity and in order to deconstruct how particular claims that contribute to a fogging effect are made and communicated, this analysis has isolated just two of the mis/disinformation narratives that circulated online after the Manchester Arena attack. The discussion has briefly referred to how they were related, inasmuch as the creation of doubt in one area opened up opportunities to seed doubts about other issues. In reality, these two falsehoods co-occurred alongside a number of others that together clouded the early view of what had transpired and why.

In the confusion and uncertainty that almost inevitably follows initiation of a terror attack, disinforming and misinforming acts can rapidly overlap and intermingle. Malignly false messages can be quite innocently transmitted and amplified by well-intentioned digital onlookers. Indeed, unintentionally misinforming messaging can be deliberately reposted even after they have been corrected, which represents one of the principal ways that extremist groups engage in ‘engineered’ fogging. This is the focus of the next section.

‘Engineered’ Fogging

Engineered fogging is about the transmission of disinformation. The analogy of a fog introduces the idea that it is not a single misleading communication, but a multiplicity of them clouding a discussion or the interpretation of an event. Thus engineered fogging involves a flurry of false claims, which can be more or less plausible, blended together such that those on the receiving end are rendered unsure what to believe or which sources to trust.

For example, during the 2013 Oslo attack, some victims used social media to contact family and friends; others tried increase their situational awareness of what was happening while they were isolated during the island shooting. However, survivors interviewed later felt that the high volume of information posted online at that time, including misinformation and attempts made by journalists to contact them while they were still in hiding, accentuated the threat that they were encountering. Another example of engineered fogging is when a group falsely claims that the perpetrator of an attack is acting on its behalf, when in reality the relationship is only that of sympathiser rather than fully fledged member. A third example is when perpetrators themselves repeat patently false claims and conspiracies, often in 'manifestos' or 'martyrdom videos', to justify and explain their violent acts.

However, the engineering of a fog that clouds public understanding is not always the product of content transmitted by those sympathetic to the perpetrator. Very often it is created by those ideologically opposed to the responsible group. A prime example of this can be observed by looking at what happened in the wake of the 2017 Westminster Bridge terror attack. On the 25 March 2017, representatives of the group Nodisinfo posted the following:

> *"It is the Zionists who are directly responsible for the staged, Islamophobic scam and hoax known as the Westminster Bridge (fake) terror attack. Regardless, whenever crisis actors are found, then, this is hard proof of a scam and hoax. Then, too, if there are producers seen on-site, along with stage hands, if, too, there are directors plus props –* ***this is absolute, undeniable proof****. Virtually all such proof can be seen in this single video…"*

Similar claims and sentiments were transmitted by adherents and affiliates of a range of hard-right and radical groups, whose social media accounts were often linked in 'small world' networks to one another, regularly posting and reposting highly polarised and ideologically loaded content.[12] Many of them invoked similar 'folk devils' in the form of Zionists and other representatives of the 'New World Order', including George Soros. Relatedly, a rhetoric of 'crisis actors', 'false flags' and 'psyops' are a recurring component of the playbook they deploy when responding to these issues. It is probably the case that such conspiratorially infused interpretations have little salience for most people, but they may play an important role in reinforcing the belief systems of adherents to the groups. Moreover, in the context of a global pandemic, the infusion of such conspiracy theory concepts into the ideologies of extremist groups with a propensity for violence has taken on increasing significance, in terms of an increasing number of anti-lockdown / anti-vaccine protest rallies, across a number of global cities that have resulted in public disorder clashes with police.

Looking across other similar incidents and the construction of deliberately misleading reactions to these, a number of other techniques of mis/disinforming and negative influencing can be distilled. These include:

- Posing questions about specific pieces of evidence rather than questioning the event overall. After Westminster, for instance, a social media user examining some of the images being circulated posted: "Isn't the lack of skid marks or blood a little odd?" This kind of invitation to an audience to participate in doing some research

[12] They would often be relatively overt in their support using particular icons, symbols and coded language in their account profiles and messages.

about what they are viewing is a technique latterly exploited to good effect by the general purpose conspiracy movement QAnon.

- Emphasising a second inflection point in terms of seeding doubt and distrust in official accounts is to highlight and accentuate any discrepancies between accounts of what happened, especially between any eyewitness details and official versions.

- A third frequently deployed technique involves attacking the credibility of other messengers. The following is an especially trenchant example:

  “Ladies and gentlemen you are being lied to, your government media emergency and security services are all involved, in fact those that many of you refer to as heroes are exactly the opposite, they are traitors to you and your country.”

4 Flooding the Zone

The techniques and the content of the messages offered in the previous section constitute the raw materials for inducing fogging. In order for this to be particularly effective, such activities have to be performed at scale. This engages a second allied concept, labelled ‘flooding the zone’, which captures how attempts are made to dominate the information space by amplifying and reposting messages in high volumes and across platforms. The dynamics of flooding are illustrated by extending the case study of the Westminster Bridge attack.

In what has become an especially infamous episode of mis/disinformation, at around 14.40 on 22 March 2017, rumours about the possible identity of the attacker emerged on social media and were rapidly amplified by news and broadcast media. However, it was soon clear that the individual identified as the perpetrator was not responsible. Despite a correction being issued relatively quickly, adherents of far-right groups used their accounts across multiple social media platforms to continue pushing this message at very high volumes as it suited the ideological narrative that they were seeking to promote.[13]

The first social media post that publicly named Abu Izzadeen, a high-profile affiliate of Al-Muhajiroun also known as Trevor Brooks, as the perpetrator, was an unverified Twitter account, @News_Executive, which tweeted:

> *BREAKING UPDATE: Reports name the Westminster terrorist suspect as hate preacher Abu Izzadeen (Trevor Brooks) from Clapton in Hackney* (@News_Executive, 22 March 2017, 17:59).

Between 2 and 6 minutes later, two foreign news outlets – *La Stampa* (Italian) and *Dreuz* (French) – also reported Izzadeen as the attacker. Both articles, however, were modified the following day to claim that Channel 4 and other British mainstream media were responsible for misidentifying the Westminster terrorist, essentially attributing the fake news to other organisations. However, even though the articles had been edited, the digital traces of social media posts allowed us to track down the original story. This was not the only instance of such editing. Fourteen minutes after @News_Executive posted, Izzadeen’s Wikipedia page was updated to claim he was responsible for the Westminster Bridge attack. There were 83 subsequent edits to the Wikipedia page that day.

Such developments notwithstanding, the false allegation against Izzadeen gained mass exposure, as intimated above, when it was

[13] Reflecting the relatively fragmented nature of the far and hard right more generally, there were multiple groupings identified engaging with these issues and narratives. Social media accounts were categorised as possessing far-right affiliations based upon assessment of the account identities and personas, their behaviour in terms of who they were following and were followed by, as well as any patterns in the content they were posting or reposting.

repeated at the start of *Channel 4 News* that evening, with the presenter stating:

> *“A source has told this programme tonight that the attacker is a man called Trevor Brooks, better known as Abu Izzadeen, a well-known member of the now disbanded British Islamist group, Al Muhajiroun. That’s news that will raise huge questions for the police and security services.”*

Of note in the above is the attribution of the story to a source to build plausibility, when presumably it was simply unverified open-source material on the internet.

Less than twenty minutes after this, Rym Momtaz, an ABC producer, tweeted that she had contacted Izzadeen’s solicitor who had confirmed that Izzadeen was still serving a prison sentence for breaching an anti-terror order and could not have been the attacker. However, while this revelation was still filtering through, both *The Independent* and IBTimes posted stories online repeating the Izzadeen allegation. Within the hour, *Channel 4 News* had issued an on air correction and Momtaz’s tweet was being disseminated in high volumes.

Some intriguing insights are revealed, however, if we track the volumes of communication on Twitter around this episode. Figure 1 was constructed by filtering the data collected based on the search terms ‘Izzadeen’ and (‘Westminster’ OR ‘#prayforlondon’ OR ‘Keith Palmer’ OR ‘Khalid Masood’ OR ‘Adrian Russell’). The Louvain Method for community detection (Blondel et al., 2008) was then run over the entire dataset to identify the top nine clusters of accounts engaging with this allegation and associated developments. The engagement of these clusters in terms of their messaging activity over time is plotted in Figure 1 below:

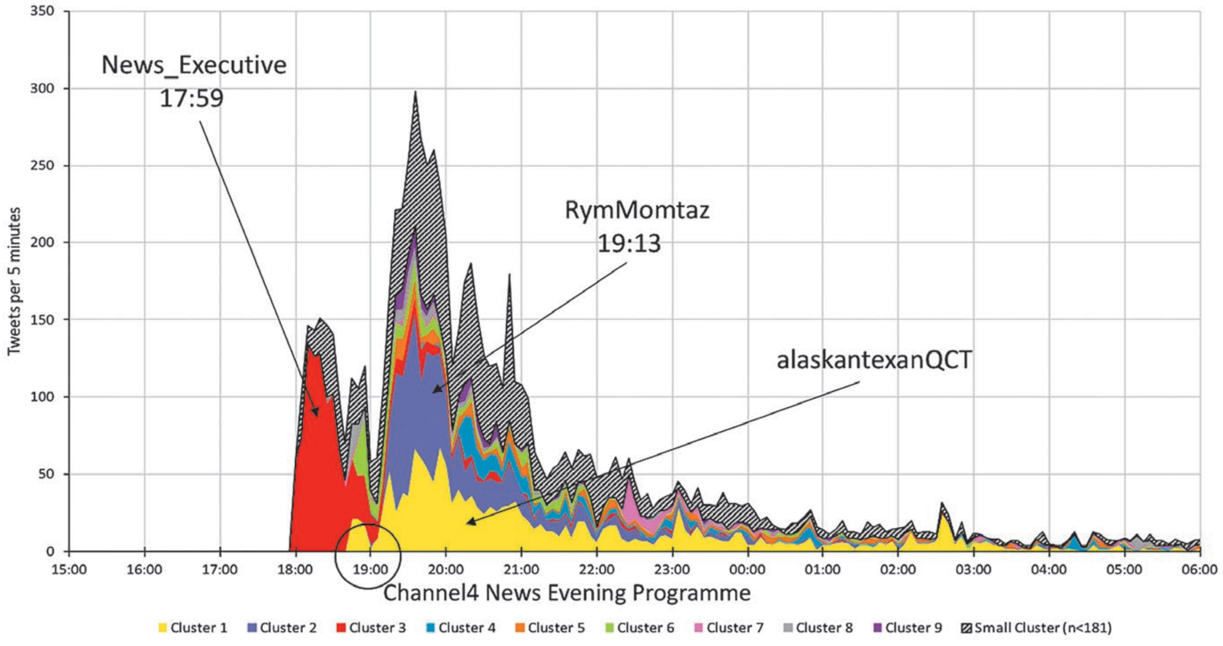

Figure 1: Temporal analysis of top 9 clusters on Twitter

In terms of interpreting this visualisation, each of the clusters is differently coloured. So, for example, retweets of @News_Executive’s original message are coloured red and dominate the initial phase. Also clearly salient is the purple cluster, depicting retweets of Momtaz’s debunking message. However, from the point of view of this paper, it is the other clusters that are of special interest. When we look at these in more detail in terms of their content, seven out of the nine clusters had their origins in posts from far-right or hard-right accounts, defined on the basis of the account identity, behaviour and content. In the graph above, one particular account cluster (coloured yellow) warrants attention on the grounds that it suggests that the messaging may not have been entirely domestic, but involved international far-right networks as well (in this case in the USA). It tweeted three provocative and ideologically loaded messages with timestamps of 18:45, 19:11 and 19:35. The first received 241 retweets, the second 1134 retweets and the third 807. In total 59 accounts, which displayed indicators of far-right and hard-right inclinations, were identified located in the USA engaging with the story.

Expressed in a different way, there were a total of 1,182 original tweets returned by the keyword searches. Of these, detailed qualitative coding suggests about 300 contained far-right sentiments, ideas or values.[14] However, filtering down to the 20 most retweeted messages, it was found that 15 of these were clearly of a far-right orientation. Moreover, these 15 were retweeted over 5,000 times, meaning that they comprised approximately 50% of all the retweets in the dataset. It is this set of dynamics that is articulated by the notion of flooding: very high levels of amplification, through both organic and bot-based means, to try and make these ideologically loaded messages the most visible.

In terms of setting out the dynamics of flooding, it is important to be clear that such messaging patterns are typically not confined to a single platform, but have a cross-platform component. For example, Jayda Fransen, a high-profile member of Britain First, used her Facebook presence, which at the time had over 259,000 followers, to post:

> *REPORTS ARE FLOODING IN THAT CONVICTED ISLAMIC TERRORIST ‘ABU IZZADEEN’ IS THE LONDON ATTACKER. HERE HE IS CONFRONTED BY BRITAIN FIRST.....*
> (22 March 17) 759 Shares.

Confirming the engagement of the global far right, the US-based American Bikers United against Jihad posted: “ABUAJ: Police shown protecting terrorist suspect Abu Izzadeen as Muslims mock patriotic British citizens.” This was also posted on 22 March 2017 and received 899 shares. The group’s total followers at the time was nearly 137,000.

Importantly, however, even when it had been established that the claim that Izzadeen was responsible was mis/disinformation, far-right and hard-right messengers on social media did not stop promoting the essence of this claim. They continued repeating and reheating it as well as variants of it for a considerable time afterwards. Obviously,

[14] Mostly this was anti-Islamic, racist and/or strong anti-immigration rhetoric. But it also included anti-Semitic conspiracy theories, and amplification of messages by known far-right influencers associated with groups such as Britain First, and individuals such as ‘Tommy Robinson’.

they had to edit and amend some details, but the fundamental ideological message was retained:

> *Abu Izzadeen represents all of the Muslim faith. Underdeveloped cult of hate and death* 🐷☪️ *#PrayForLondon…*
>
> *Not Abu Izzadeen, but still a Muslim radical – who would’ve guessed? London attacker named as Khalid Masood.*

An important insight here is that the adherents of extremist groups who expressed these kinds of sentiment stayed engaged with this event for a much longer period of time than the majority of social media users. There was a clear spike in attention for topics and hashtags connected with the Westminster Bridge attack for several days, but then the online public chatter largely moved on to other subjects. However, far-right extremist chatter persisted in its fixation on the kinds of ideas featured in the messages reproduced above. This may be a secondary consequence of flooding the zone.

Flooding thus represents an informational effect distinct from fogging. The latter constitutes the general condition that is induced by the co-occurrence of multiple distorting and misleading messages that creates a conducive environment for its conceptual cousin. Flooding involves a surge of communicative activity around a particular message and variants of it, such that it contributes to the reproduction and reinforcement of the fog.

5 Policy and Practice Implications

For understandable reasons, whenever a terrorist attack occurs, the main focus for the police, intelligence services and allied agencies is on managing the material threats and harms involved. That said, there is a growing recognition that such situations represent a specific and important strategic communications challenge, given how what is communicated across social media can have a profound effect in shaping how the situation is defined and its longer-term consequences. Rumours and conspiracies sown in the aftermath of a tragic event can prove difficult to counteract over the longer term if they gain traction.

Framed in this way, the preceding discussion has several implications for policy and practice:

- First, there is the conceptual connection of linking disinformation with our understanding of social and political reactions to terror attacks. Notably, disinformation has tended to be linked to problems of election interference and similar democratic events, rather than post-terrorist violence. But as has been demonstrated above, disinforming, distorting and deceptive messaging can play a particular role in shaping the trajectory of practical responses on the ground and wider public understanding in the wake of an attack.

- Having evidenced that disinformation arises and proliferates in the post-attack environment, the second real world implication is that the key concepts suggest that there are patterns to the manipulated and manipulative information being transmitted. As such, there is potential for algorithmic interventions to reduce the frequency and/or visibility of any such malign information, thereby contributing to the reduction of the overarching public harms induced following terror incidents. This may be especially the case in terms of using automated methods to detect and reduce the impact of ‘flooding’ the information zone in the aftermath of violence, given this typically pivots around the high frequency reposting of variations of a key message or image.

- It is worth noting in the case examples discussed above that there were multiple, competing sources of disinforming and misinforming content. For example, it was not simply coming from the ideological supporters of an assailant or merely from their opponents. There were instead sequences of claims and counter-claims communicated with a variety of intentions. Thus, the challenge for policy and practice is to identify interventions that are ‘actor agnostic’ and reflect the increasingly multi-polar make-up of media audiences, where there are typically multiple influence vectors simultaneously at play.

6 Conclusion

This report has introduced three concepts to help to understand how and why mis/disinformation arises in the aftermath of terror attacks and some of the implications and consequences that can be attributed to it when it does. It has been suggested that some misleading messaging is purposive and constructed via the deployment of particular techniques of disinformation. While similar forms arise more organically, they can nevertheless be harnessed and manipulated for malign purposes by actors motivated to do so.

More broadly, this report demonstrates the value of further dialogue between researchers engaged in the study of different forms of online harm. Specifically, the intent herein has been to appropriate some of the analytic frames used in the rapidly growing sub-discipline of disinformation studies, which has most often been applied to electoral processes, to show how such frames also afford new insights and understandings in relation to the evolution of public opinion and understanding following terror attacks. A key finding is to show how consequential and influential communications of this kind are not just made by those ideologically aligned with the violent actor, but frequently flow from the responses and reactions of their ideological opponents. Thus, the analysis points to some of the complexities and nuances involved in managing and mitigating the harms that arise in the aftermaths of terrorism.

Policy Section

This policy section has been written by Inga Kristina Trauthig, research fellow, and Amarnath Amarasingam, senior research fellow, at the International Centre for the Study of Radicalisation (ICSR) at King’s College London.

The key findings of this report carry corresponding policy implications for governments around the world. At the same time, technology companies are well aware that they are at the forefront of addressing communication on social media in the aftermath of a (potential) terrorist attack. The following section seeks to achieve a threefold aim: first, to deliver concrete policy recommendations for governmental stakeholders; second, to outline policy options and strategic foresight for technology companies; and, finally, and in hand with [1] and [2], to serve as a reference point for future evaluation of tech policies in order to assess dos and don’ts of technology legislation around the globe.

With this, the policy section ensures that the Global Network on Extremism and Technology (GNET), the academic research arm of the Global Internet Forum to Counter Terrorism (GIFCT), is academically advising and supporting technology companies and policymakers on how to better understand the ways in which terrorists are using information technology. This is designed to fulfil not only GIFCT’s pillar of learning, but ultimately to improve prevention and responses to terrorist and violent extremist attacks.

1. Focus: Policymakers

This report’s examination of the drafting and amplification of misinformation and disinformation in the immediate aftermath of terrorist attacks carries relevant implications for national and international (EU, UN etc.) policymakers, especially homeland security officials ranging from law enforcement to social workers engaged in prevention programming.

- To start with and given the speed in which misinformation and disinformation develops and spreads after terrorist attacks, an **impactful addressing of this harmful dynamic requires fast-tracked cooperation between governments, in particular law enforcement, and technology companies**. In order to avoid confusion among technology companies about which content to allow and which to moderate, ban, or label as misleading, because it could potentially cause public harm or is deliberate disinformation, these companies would benefit from being able to rely on the most solid information attainable at that given moment in time. This information is likely to be held by law enforcement called to the scene rather than open-source information that this research has shown to be targeted, tainted and flooded in the aftermath of an attack by sympathisers, adversaries and bystanders.

- For a holistic government approach, lawmakers could address a three-pronged understanding of terrorists’ communicative actions. First, enable regulations that target terrorist communication, such as livestreams – for example, in cases where perpetrator livestreaming is not legally codified as forbidden this should be done in order to offer tech companies a more straightforward mandate to act in such an instance; second, facilitate countermeasures to disarm the burgeoning framing contest – this could include efforts by law enforcement to directly engage with social media users and posts which state misleading claims in the immediate aftermath of an attack; third and related, **invest resources in social control responses, including social media messaging by governments, police and intelligence agencies**, in terms of both the practicalities of any investigation and public reassurance. The speed with which accurate information from trusted sources can be released to the public is fundamentally important for ensuring that misinformation does not spread unimpeded.

- Since this GNET report has examined how fogging, flooding and surfacing pollute the post-attack information space, **law enforcement dealing with the aftermath of an attack is well advised to aim for transparency and coherence** as much as possible. The outlined framing contest in which ideological supporters or opponents of the perpetrator try to establish a public definition of a given situation in terms of how the violence is interpreted and understood illustrates the relevance of this.

2. Focus: Technology Companies

Next to the stringent commitment to support the outlined cooperation between governments, law enforcement and technology companies, certain steps can be taken by technology companies without relying on lawmakers’ backing.

- Similar to measures that Facebook undertook in the immediate aftermath of the 2019 Christchurch shootings in New Zealand, societies benefit from technology companies being committed to pinpointing and deleting videos depicting violent attacks made by the perpetrator or an accomplice, which are either livestreamed or uploaded after the attack. In line with the existing GIFCT incident protocol,[15] ensuring that the same content cannot be uploaded across platforms, in its original form or in edited versions, is also important. Those account holders who are dedicated to uploading videos of massacres for their own political or ideological purposes should face substantive repercussions, including having their accounts suspended. For depictions of violence generally, tech companies should work to build repositories accessible for researchers or international human rights lawyers for instance, working on the documentation and analysis of war crimes.

- As the report has shown, disinformation arises and proliferates in a post-attack environment, and manipulated and manipulative information can often be transmitted. Therefore, technology

[15] “Content Incident Protocol”, Global Internet Forum to Counter Terrorism, https://gifct.org/content-incident-protocol.

companies and (internal as well as external) researchers could continue to work side by side to gauge if **algorithmic interventions** could reduce the frequency and/or visibility of any such malign information, and if so, which interventions. In this fashion technology companies would contribute to the reduction of the overarching public harms induced following terror incidents that this report referenced.

- Social media companies could increase transparency with regard to a clear, comprehensive communication plan for emergencies. This GNET report has argued that misinformation, disinformation and hate-speech circulate after a violent attack. Therefore, measures could be put in place to limit the dissemination of these kinds of content. This could include encouraging sharers to read the content before disseminating, flagging or removing unverified content that is going viral, or limiting forwarding behaviour for short periods of time after an attack.

- Finally, this GNET report has outlined that misinformation not only fans out on social media but also traditional or broadcast media. In line with this, **technology companies could continue to develop open channels with those in traditional media to share limited information about what technology companies see on the backend, so that journalists do not unwittingly amplify unverified content**. This could be along the lines of what GIFCT members, such as YouTube, already do within its News Initiatives, for instance; these initiatives work to have technology firms become partners with rigorous journalism and a healthy digital news ecosystem.

3. Focus: Strategic Foresight and Broader Implications

Next to the policy recommendation derived directly from the quoted GNET report, broader implications and strategic deliberations can be retrieved from this study of the disinformation development and spread in the wake of terrorist and violent extremist attacks.

- Since this GNET report focused on open-source data and hence information that is publicly available, it disregarded communication that spreads in closed-off spaces such as **messaging apps like WhatsApp or groups and forums that you have to request to join, found on a wide variety of platforms, such as VK**. For future research, the inclusion of these spaces would be valuable in order to assess similarities and differences in the content creation and spread of disinformation in the wake of terrorist attacks. Access to these groups is more difficult but possible, however ethical concerns always accompany this sort of research. Therefore, academics, policy makers and tech companies are well-advised to constantly re-evaluate their policies and practices to facilitate legal ways for research but also protect privacy of users, who often use these spaces to escape state surveillance in authoritarian countries. Subsequently, these research insights would further inform ongoing discussions among policymakers and technology companies that relate to regulatory deliberations with regard to limits on group chats and actions such as (limiting/marking) automatic forwarding on **platforms and forums that are more difficult to moderate**, such as end-to-end encrypted (E2EE) messaging apps or invite-only forums.

- Another related policy aspect that has gained traction over recent years is focusing attention on the victims of terrorist attacks. Since the report outlined the distress experienced by many victims in the online space about the high volume of information posted online at that time, including misinformation and attempts made by journalists to contact them while they were still in hiding, **policies that focus on protecting the survivors of terrorist attacks should take into account the online space** and cooperate with tech companies in how best to shelter victims in the wake of an attack as well as the following months.

Global Network
on Extremism & Technology

CONTACT DETAILS

For questions, queries and additional copies of this report, please contact:

ICSR
King's College London
Strand
London WC2R 2LS
United Kingdom

T. **+44 20 7848 2098**
E. **mail@gnet-research.org**

Twitter: **@GNET_research**

Like all other GNET publications, this report can be downloaded free of charge from the GNET website at www.gnet-research.org.