A framework for misinformation crises

Paper 1: Identifying scope and risk

18 November 2020

Purpose

We know that certain events can affect the information environment by prompting an increasing complexity of accurate information, confusion, or by creating information gaps - all of which can result in an increase in the volume of misinformation. This was clearly evident during the coronavirus pandemic, which prompted a slew of intensified counter-misinformation measures from internet companies, governments, media, fact checkers, academics and civil society.

The response to coronavirus misinformation this year has shown how fast and innovatively those working to analyse and counter it can respond. But it has also thrown light on the need for greater discussion of principles, proportionality, and the use of evidence in responding to other types of future information incidents that may be just around the corner.

That is why Full Fact is bringing together practitioners, experts and community groups from different sectors affected by and aiming to affect the information environment to develop a framework to identify the issues that occur during moments of crisis and develop joint aims for how organisations should respond. Our aim is to develop something simple and useful that can help specialists in this area coordinate our work, and outside stakeholders understand it.

A set of papers will inform the development of the framework. These papers will be published on the Full Fact website to help stakeholders explore issues in greater depth which are raised during outreach with experts and to enable further discussion and debate as we build towards a common understanding for tackling these problems in the future. Many thanks to Facebook for funding this project, and to all the experts who have contributed so far.

Outline

This paper (#1) sets out the first two issues we are exploring for the framework: identifying the types of incidents likely to be in scope; and outlining an initial classification of factors that could indicate risk or severity of an incident.

Developing these two categorisations will be the foundation for building a framework that targets the right type of situations and understands the types of responses that may be appropriate. Feedback on how these two elements could be refined further is welcomed (please see the end of this document on how to provide input). Full Fact is grateful to everyone who has provided feedback on this paper to date.

Incidents in scope

Our aim is to identify the types of incidents that have or are likely to have a substantial and material impact on the way information is consumed or shared by the general public.

We want to understand which crisis incidents have potential for a member of the public to come across mis/disinformation. It may be that something unexpected and unexplained has happened, like a terrorist attack, and people are actively looking for information to understand more. Or there could be a polarising event coming up, such as an election, and people are coming across information that is being pushed out. In both these scenarios the baseline information environment shifts in some way, and information might be new, complex, or confused.

Based on a comprehensive mapping of recent such incidents, we have developed nine categories that we consider would often require a counter-misinformation response and which would fall in scope of the framework. These groupings attempt to bring together types of events that have similar significance. Please note that these are in no particular order.

Not all of these incidents would call for the activation of the framework - later we describe indicators for understanding severity, which can help distinguish what types of incidents might merit a response over and above day-to-day counter-misinformation efforts.

Categories of incidents

1. Human rights or freedom of expression abuse:
Peaceful civil action such as protests; violent public confrontation; long term escalating tension between regions; mass detainment and killings; citizenship or demographic changes.

2. Unexpected disasters with high, wide reaching impact:
Deliberate attack with wide reaching, long term impact; innovative or novel deliberate attack; incidents that create deaths or displacement of people; national/regional pandemic.

3. Unexpected events but where there is a level of preplanning and low or short term impact
High impact weather; incidents with cultural/religious significance; localised deliberate attacks; accidents and transport disasters; hack and leak or data dumps. Authorities would be expected to have a plan for response or containment**.**

4. Long-horizon and long-tail incidents:
Economic changes; food or fuel shortages; exceptional electoral events; displacement of people; scientific developments.

5. Democratic, planned political events (versus undemocratic elections)
National or regional votes.

6. Politically and/or emotionally sensitive anniversaries of nationally-significant events where there could be opportunities to exploit polarisation:
War memorials; societal or political commemorations; controversial incidents.

7. High impact incidents that occur or where the impact is felt across multiple markets:
Pandemic; outbreak of war.

8. Controversial and/or shocking news stories:
Events that generate news headlines, but do not become major incidents.

9. Spreading of false/misleading information by authoritative actors
Some party political messages; high profile endorsements of conspiracies; denying controversial incidents.

This categorisation deliberately does not consider consistent “low level” streams of mis/disinformation. At this point we are considering the acute rather than the

chronic, but we hope this framework builds a foundation on which to better understand misinformation that is prevalent in everyday life.

Nonetheless, it is helpful to consider when an event may tip over into becoming an incident in scope of this framework. We suggest criteria for that below.

Determining risk

Judging the "harm" or "impact" caused by bad information is a notoriously difficult subject, and no consensus has emerged on an effective way to judge this. Instead the trend is to use a variety of proxies such as reach to judge what level of harm could have been caused.

The eight categories below attempt to set out a comprehensive category of questions to consider when judging both the initial risk that an incident may impact on the information environment, and the subsequent severity of an incident. From this it may be possible to consider the most appropriate action depending on the level an incident is at, or could be moving towards.

Indicators to determine risk	
1. Gravity	What is the likelihood of immediate or lasting real world harm from bad information (for example, violence, physical danger, loss of herd immunity, democratic inefficacy)?
2. Scale	How many people is this reaching? Is the reach growing? Is it growing beyond audiences who have similar world viewpoints to reach new groups?
3. Novelty	Has the primary audience experienced similar incidents before? Does this play into existing misinformation narratives/conspiracies?
4. Timeline	How long is this incident likely to last?
5. Predictability	Is there time to put plans in place in advance?

6. Cause	Is there clear malicious intent, is there one actor driving misinformation or is there clearly someone to blame?
7. Polarisation	Does this play into existing social or cultural controversial or polarised topics and ideologies?
8. Demographics	Are certain vulnerable or otherwise marginalised groups being targeted with or by information?

In future papers we will look to match up responses to this categorisation with what the aims should be in response, and potential best practices to achieve that aim.

It is our intention that these questions could be applied at any point in an incident's lifespan, and aims and responses agreed either in advance or reactively:

- When an incident is just emerging, these could be helpful in assessing whether responses are needed in advance or if preparation should begin
- During an incident these could be used to judge which actions are likely to be most effective in response
- Once an incident is complete, these could be used to evaluate responses and whether the right judgements were made at the right time

These questions are not currently in order of prioritisation, but we recognise this will be important in determining an overarching assessment of risk and therefore severity. This will also be reviewed going forward in the project.

It is also unlikely that any one organisation can provide an answer to all of these questions. This is to be expected and in our view would not hinder an initial assessment. Not having a clear answer to any question could, in itself, point towards certain responses or encourage cooperation between organisations. Considering uncertainty explicitly will help make the framework a practical tool.

Next steps

We are publishing this paper to encourage wider feedback on how our thinking is developing. We hope that these two categorisations are helpful in understanding the scope of the project, as well as how to identify events that may require a greater or lesser response. We would be interested to hear thoughts from other organisations who are involved in responding to bad information on whether:

- The right events are identified, or if there is anything missing;
- The indicators are comprehensive and if there are specific priorities within these indicators.

Building on this foundation, we will next explore what the common problems caused by the incidents identified are likely to be, outline where there are commonalities or lessons that can be applied across incidents and conversely where there may be specific considerations that only apply to certain incidents.

Please do get in touch if you have any feedback on this paper, or would like to contribute to this work, at phoebe@fullfact.org. Please note we may be unable to respond to every contribution.