Urgent: A Structured Response to Misinformation as Harm

Working Paper
September 2022

Connie Moon Sehat,
Tarunima Prabhakar,
Ryan Li,
Peipei Nie,
Amy X. Zhang

Social Futures Lab,
University of Washington,
in collaboration with
Hacks/Hackers

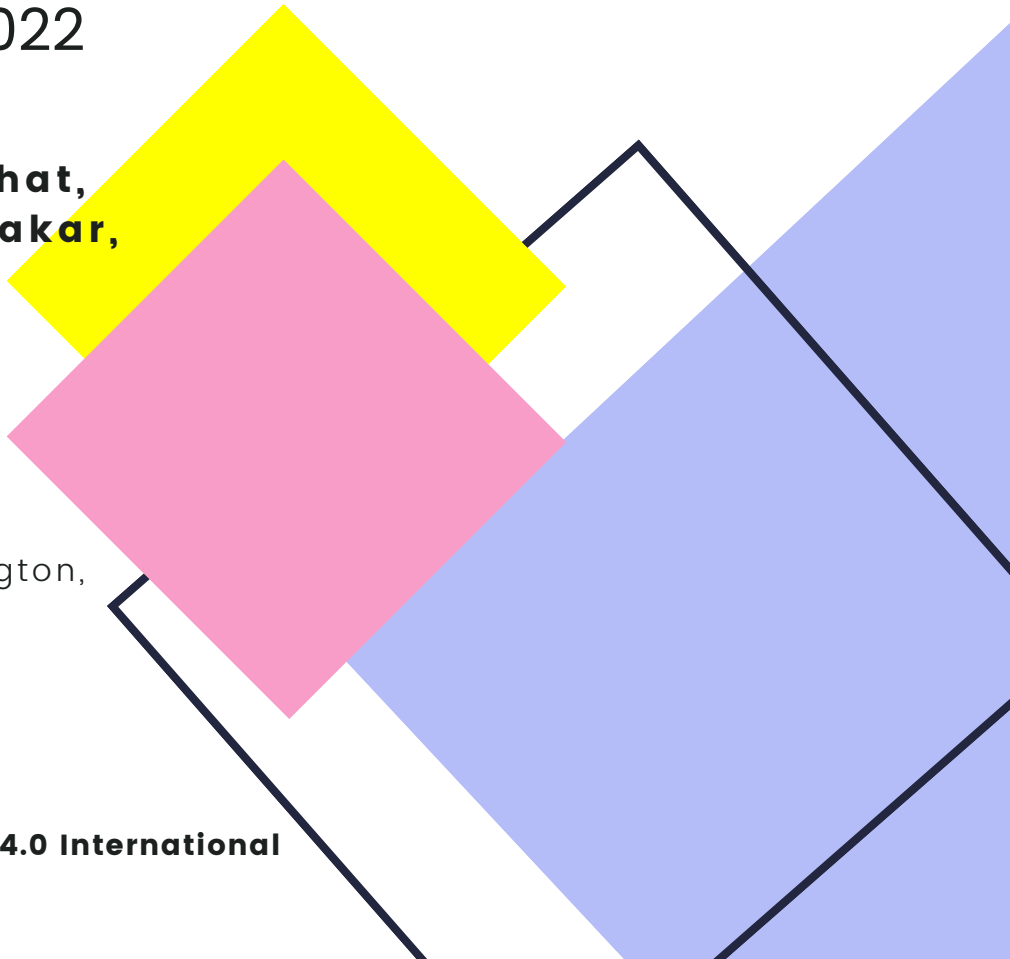
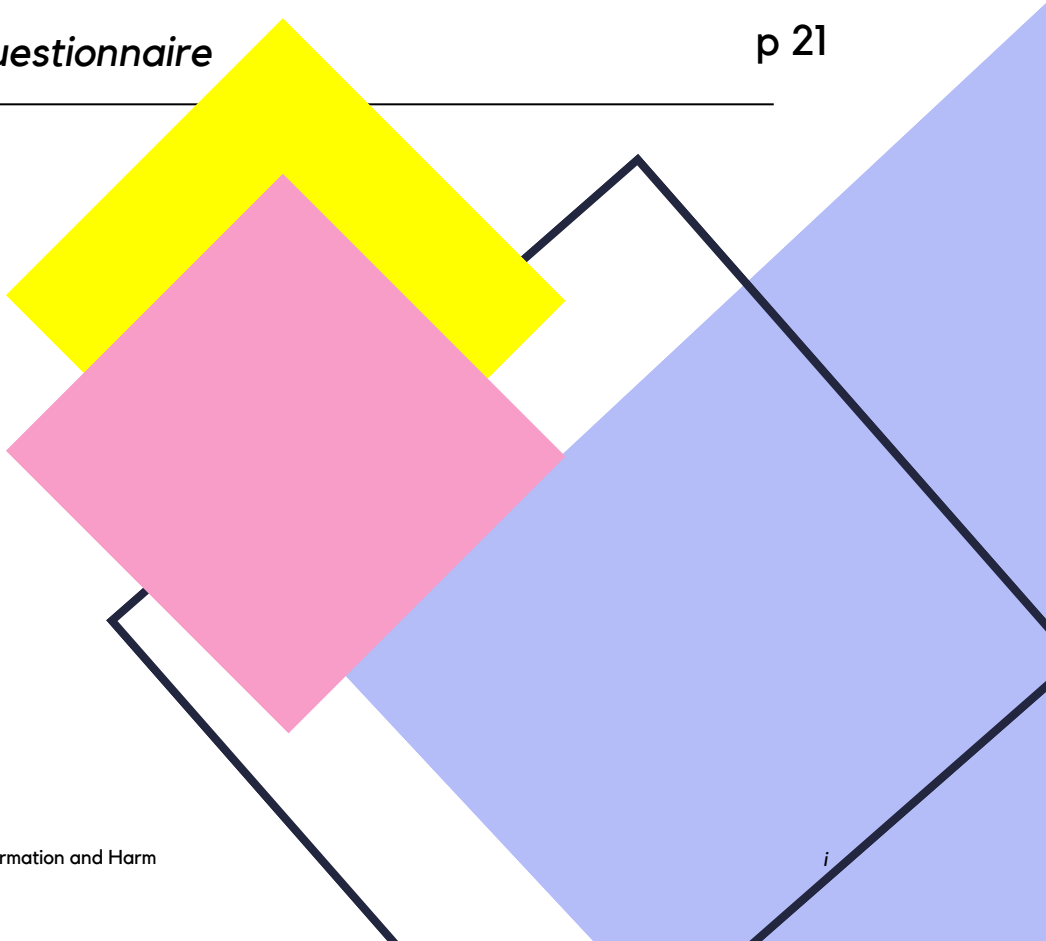


Table of Contents



Executive Summary

Online misinformation is a major challenge for societies today. Beliefs in false claims about science, such as vaccine misinformation, can lead people to engage in harmful behavior that risks their own health. Such misinformed beliefs can also defeat public health measures that rely on collective compliance to protect society's most vulnerable. Similarly, a belief in inaccurate or misleading narratives about topics such as vote-rigging or other supposed election interference can lower the public's trust in democratic institutions, and in turn affect the level of participation in political activities such as voting, interfere with the peaceful transition of power, and even motivate political violence.

Fact-checking is a critical activity when addressing misinformation. Fact-checking supports individual readers who seek good information, and also supports content moderation initiatives on larger scale platforms. However, fact-checking is laborious. The fact-checking process includes investigating claims, collecting convincing evidence that such claims are false or misleading, and then sharing that evidence out. With torrential volumes of user-generated content being created daily, it is impossible to fact-check every new article, post, message, or claim.

As a result, fact-checkers tasked with addressing online misinformation must prioritize what they choose to tackle every day. Given that prioritization is unavoidable, how should fact-checking efforts to combat misinformation prioritize what content to tackle? A working group of academics, non-governmental organizational researchers, and students based in the Social Futures Lab at the University of Washington's Allen School of Computer Science and Engineering decided to explore this question.

From interviews that the writers of this paper conducted with fact-checkers, we found that fact-checking processes are still young and not standardized as a field. Fact-checkers typically take a relatively ad hoc approach to prioritization, using individual judgment and case-by-base discussion with others. Could prioritization instead be achieved in a principled and systematic way? One way forward that we propose is via **harm assessment**.

In applying a structured harm assessment to misinformation, we begin by making the observation that while all misinformation is harmful to

some degree, not all misinformation is equally harmful. Following a literature review, and a series of interviews and workshops with fact-checkers and other misinformation experts, we identified major dimensions for assessment.

Five dimensions — **actionability, exploitativeness, likelihood of spread, believability, and social fragmentation** — can help determine the potential urgency of a specific message or post when considering misinformation as harm. In addition, we conclude this paper by providing a checklist of questions to help determine a piece of content's relative level of urgency within each dimension.

The dimensions and the questionnaire are intended as both conceptual and practical tools to support fact-checkers, content moderators, peer correction efforts, and other initiatives as they make strategic decisions when prioritizing their efforts to respond to misinformation that is spreading.

Background

Fact-checking is a critically needed field that is evolving to meet the scale of online exchange

It's clear that the facts matter: Misinformation and disinformation around the COVID-19 pandemic or elections around the world can have impacts ranging from peoples' health to the stability of democratic institutions. Because of the potential impact and speed at which content can spread online, fact-checkers are often asked to work quickly. In the wake of the January 6 attack on the US Capitol, for example, TikTok noted that its coalition of fact-checkers could return results within the day, if not hour.[1] Yet even this speed is not enough, given the nearly instantaneous and massive reach that false rumors seem to enjoy.[2]

Faced with this challenge, fact-checking has entered a new era. Online platforms that focus on user-generated content have been increasingly exploring ways to scale content review, in order to match the pace of online distribution, as well as to support safe and accountable exchanges of information. This is one reason for an explosion of growth in the field of fact-checking, which originally grew out of magazine journalism in the earlier half of the twentieth century.[3] Fact-checking organizations now form independently from journalism outlets — for example, a recent census counted 391 fact-checking organizations in 2022 in contrast to 186 in 2016.[4] In addition, the demand for accurate information in elections and health contexts itself has changed the nature of fact-checking work.[5]

Despite such demand, current-day fact-checking is a field that has its critics. The complicated issues regarding the line between fact and fiction itself cannot be captured by employing a simple "true/false" rating.[6] Fact-checkers can also be biased in ways that affect which facts matter to them, and how.[7] Their individual decisions about which information to address is especially important to recognize within democratic societies, where equal voices and participation in public discussion is desired. However, criticism notwithstanding, fact-checking efforts have demonstrated an ability to agree across organizations and gain greater consistency in their assessments.[8] There is promise, then, in

[1] Perez, "TikTok to Flag and Downrank 'Unsubstantiated' Claims Fact Checkers Can't Verify."

[2] Vosoughi, Roy, and Aral, "The Spread of True and False News Online."

[3] Graves, Deciding What's True; Currie Sivek and Bloyd-Peshkin, "Where Do Facts Matter?"

[4] Stencel, Ryan, and Luther, "Fact-Checkers Extend Their Global Reach with 391 Outlets, but Growth Has Slowed."

[5] Fischer, "Fact-Checking Goes Mainstream in Trump Era"; Siwakoti et al., "How COVID Drove the Evolution of Fact-Checking."

[6] Bell, "The Fact-Check Industry."

[7] Ceci and Williams, "[OPINION] The Psychology of Fact-Checking."

[8] Amazeen, "Revisiting the Epistemology of Fact-Checking."

making fact-checking processes more systematic and impartial, for the benefit of society.

As fact-checkers are presented with hundreds of requests each day for claims to evaluate, formalizing a triage approach may minimize biased decisions by individual fact checkers, as well as provide a common vocabulary that invites others to discuss the effectiveness of these efforts. In addition, deciding what factors should determine triage is tantamount to deciding what is most important for fact-checkers to address, or where fact-checkers can have the most impact. By making these factors legible through a structured and documented process, fact-checking groups also clarify their principles to themselves, to each other, and to the broader public.

Prioritization is a critical need in particular.

We ask:

How can fact-checkers be supported in their decision making to decide which pieces of content may have more negative impacts in comparison to others?

To answer this question, we observed that prioritization might be improved by **considering false claims, or misinformation and disinformation, as potential harms in themselves**.

Our approach: Address misinformation as a harm

The question of whether false information is harmful itself is perhaps as old as human society. Do all untruths damage others? What if they are intended to prevent harm, such as white lies? How about misleading statements or omissions of fact? Philosophers and theologians have certainly been engaged in questions around truth and falsehood for millennia. For the last two decades, concerns about whether we are in a "post-truth" society from journalists, political scientists, cognitive psychologists, and other research communities have added to this conversation. Even as finer points are debated, scholars acknowledge the deleterious effects of lying upon interpersonal trust, overall sociability, and even the ability to hope.[9]

Taking the approach that misinformation could be treated as a harm opened up a fruitful line of inquiry. The perspective of misinformation as a harm aligns with the motivations of fact-checkers. Like the journalism field out of which it was born, fact-checking has at its heart altruistic ideals such as holding power accountable

[9] Providing an overview of the moral dilemmas behind lying at an individual and social level, even the "harmless white lie," is Bok, *Lying*; Regarding the concept of post-truth, an early example can be found in Keyes, *The Post-Truth Era*; more recently, see also Lewandowsky, Ecker, and Cook, "Beyond Misinformation"; Franklin and McNair, *Fake News*; Iyengar and Massey, "Scientific Communication in a Post-Truth Society"; Farkas and Schou, *Post-Truth, Fake News and Democracy*; Regarding the impact of falsehood upon hope, see Snyder, "Hope Theory: Rainbows in the Mind."

and helping the public to achieve informed decision-making.[10]

Practically speaking, existing moderation practices for addressing potentially harmful content face the similar challenge of triage. Content moderators also must contend with a tidal wave of content via the Internet. Even with automated AI tools to help remove spam, platforms such as Facebook still had over 3 million pieces of flagged content every day in 2020.[11] Reports include people working to review between 25 to 100 pieces of content every hour.[12] More examples of this scale and challenge can be seen in reports from other platforms such as YouTube or Reddit.[13]

Moreover, harm itself is a complex social and legal concept that involves a process of clarifying, or classifying, its relative degrees of effect and corresponding proscriptions or punishments.[14] In online realms, harm characterizes a wide range of socially undesirable online content, such as harassment, abuse, and child exploitation. These definitions of harm, like those related to truth, can be socially dependent and may involve the evaluation of multiple incidents over time, making context and nuance critical. And, at least within democracies, attempts to assess relative degrees of harm must maintain the fine balance against diminishing other human rights regarding the freedom of speech and conscience and rights to free assembly.[15]

Bringing a structured harm assessment to misinformation, then, means to prioritize according to its potential harmful effect: **While all misinformation is harmful to some degree, not all misinformation is equally harmful**. As an example, compare a hoax about a celebrity death versus the false claim about toxic seeds that supposedly provide COVID-19 immunity.[16] For a variety of reasons, it may be more urgent to try to combat the latter example of misinformation, before worrying about the former.

Not only can articulating the *relative harmfulness* of a piece of misinformation help with prioritization, but such articulation can also help with measurement efforts. Organizations that seek to measure the volume of misinformation spreading on a platform or determine the impact of an intervention to reduce misinformation spread may be more interested to focus their measurement on *harmful misinformation*, according to the degree of harm that they prefer to measure, as opposed to all forms of misinformation.

---

[10] Graves, Nyhan, and Reifler, "Understanding Innovations in Journalistic Practice."

[11] Barrett, "Who Moderates the Social Media Giants?"

[12] Ibid.; Shead, "TikTok Is Luring Facebook Moderators to Fill New Trust and Safety Hubs."

[13] "YouTube Community Guidelines Enforcement – Google Transparency Report"; "Transparency Report 2021 - Reddit."

[14] Feinberg, The Moral Limits of the Criminal Law Volume 1.

[15] For an overview of rights-related balances and tradeoffs when it comes to harm, see Sehat and Lalani, "Advancing Digital Safety: A Framework to Align Global Action."

[16] See for example https://en.wikipedia.org/wiki/Death_hoax; in contrast, "Twelve Taken Ill after Consuming 'Coronavirus Shaped' Datura Seeds."

Looking to other approaches to address harm and misinformation

Both misinformation and harm are complex concepts to address. In order to develop our approach, we looked first to researchers who in the last several years have started to develop frameworks and definitions regarding harm and misinformation; these frames and definitions support identification and prioritization. We then aimed to distill these works into categories that could support our efforts to address misinformation as a harm.

Some examples of work around taxonomies and frameworks that we found inspiring include:

- Agrafiotis et al. 2016 and 2018 defined a taxonomy of harms from a cybersecurity perspective. Dimensions of harm defined include economic, social, reputational harms.[17]
- Scheuerman et al. 2021 established a framework of severity for harmful online content, by considering approaches of severity from legal, law enforcement, and health professional perspectives. Taking 66 categories originally from Facebook, the authors further refined the classifications to 20, with misinformation appearing within the single category of Coordinating Scams and Political Attacks.[18]
- FullFact white papers in recent years have attempted to establish a framework of severity around "information incidents," or large-scale public incidents where the coordination across organizations and institutions may be helpful. This approach is less about incidents regarding individuals (unless public figures), and more concerned around issues such as public health and elections; misinformation is clearly of concern in these issues.[19]
- Categories of misinformation identified by First Draft and Wardle and Derakshan have offered many good starting points. The main focus of these conceptual definitions is false and misleading information though the additional category of 'malinformation', or factual information used to inflict harm, does clearly intersect harms-related work.[20]

[17] Agrafiotis et al., "Cyber Harm"; Agrafiotis et al., "A Taxonomy of Cyber-Harms."

[18] Scheuerman et al., "A Framework of Severity for Harmful Content Online."

[19] For example, FullFact, "Towards a Framework for Information Incidents - Paper 3: Levels of Incidents."

[20] Wardle and Derakhshan, "Information Disorder."

Reading through the above works:

- Certain characteristics define different general types or *categories* of harm.
- At the same time, other characteristics capture *variable magnitudes* of harm whose value may vary according to context.

For example, categories of harm may be physical versus economic harm, or harm that affects an individual versus society at large. A recent work that noted potential "misinformation harms," or harms arising out of misinformation, also made use of categories to distinguish different types of harm.[21] An example of a characteristic of harm that is variable includes the reach of a piece of misinformation, the value of which can change to be higher or lower depending on factors such as the popularity of the poster or the platform where information is being shared.

Our contribution: a model of variable "dimensions of urgency"

When characterizing misinformation as a harm in order to assess urgency, we have found it useful to clarify the difference between categorical and variable characteristics, as **we can more easily define degrees of urgency within a category of harm as compared to across categories.**

For example, harms that are more directed at individuals, such as doxxing or exploitation, are different from those affecting broader society, such as health or election misinformation. How does one "calculate" that the first kind of harm is less impactful than the second? What happens, for example, when that first category becomes an issue of child sexual exploitation? When it comes to policies for handling certain categories of misinformation, the appropriate range of actions is likely to be pre-defined through a mix of research and community or expert consensus, rather than through any dynamic variables or individual judgment, on the fly. Yet within categories, it may be possible to clarify which cases might be more urgent to handle in contrast to others. However, we note that categorical considerations still need to be accounted for on some occasions. For example, imminent physical harm has consistently been an important threshold regarding urgent response.

Using an iterative approach, we created a questionnaire that aimed to isolate distinct dimensions of misinformation as harm that can signal a degree of urgency. We incorporated thematic concepts for evaluating harm that we discovered in our review of existing literature, as well as insights derived from past engagement with those working in fact-checking. Through internal sessions and in workshops with other experts, we further distilled these dimensions as we refined the questionnaire.

At the same time, we interviewed 23 fact checkers from 15 countries to better understand their thoughts around the harmfulness of misinformation and their work process to validate the approach. Qualitatively "coding" their responses—or

[21] Tran et al., "Misinformation Harms."

categorizing answers according to themes—also enabled us to amplify the framework, resulting in additional questions and the creation of a dimension that we call "social fragmentation," which addresses longer-term community or societal level harms. The fuller description of these interviews and work is being developed for peer-reviewed publication.

The end result is a framework of five variable "dimensions of urgency" that can support fact checkers in their efforts to discuss and prioritize in a more strategic way. These five dimensions can help to clarify urgency within a specific category of misinformation as well as harm, thereby helping response teams. By providing multiple dimensions, we aim to offer a holistic approach that encourages the evaluation of issues that might be missed if organizations are always only focused on responding to the most urgent content. The five dimensions can also be further developed in future research around what is possible to automate in practical interventions.

We note that our framework assumes that degree of harm is positively correlated with degree of urgency. However, this assumption may not always hold, particularly for some organizations for whom urgency might also involve other factors, such as internal resources. In those cases, organizations can still use our framework but adjust it to their needs.

Our misinformation prioritization model: Five dimensions of urgency

Our model includes five dimensions of urgency when it comes to assessing and prioritizing misinformation as a harm:

1. Actionability
2. Exploitativeness
3. Likelihood of spread
4. Believability
5. Social fragmentation

The five dimensions of urgency are each defined through a set of questions, which are incorporated into a single questionnaire (see Appendix); the questions that make up the questionnaire reflect factors that are currently understood to have an impact upon misinformation's magnitude or potential harm.

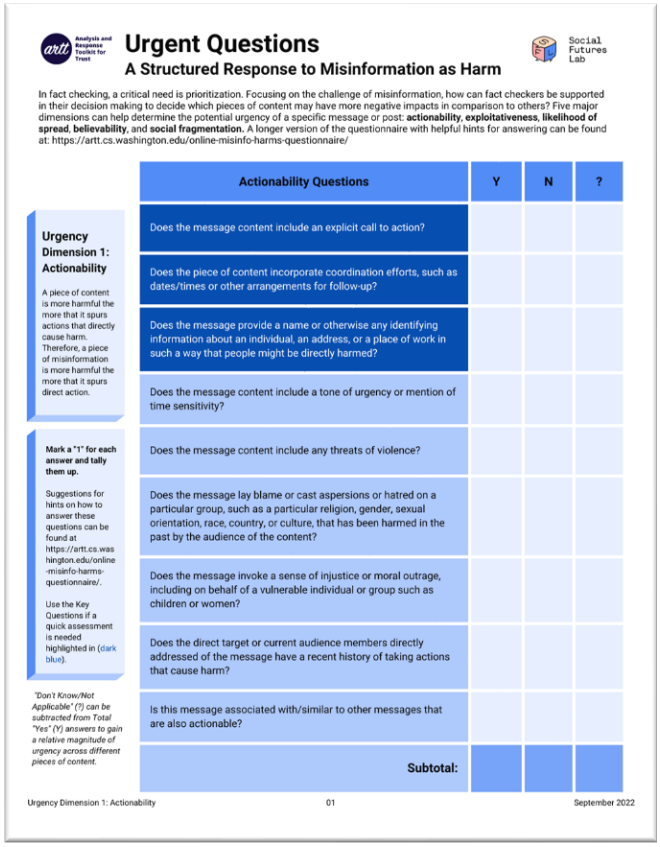
Some aspects of the questions are content-agnostic, such as whether a "call to action" takes place. Other questions are both content and context dependent. For example, directly mentioning vulnerable populations that have been the target of misinformation campaigns. While we interviewed fact-checkers working in multiple languages, these questions do lean upon current understandings grounded in primarily English-language research. Questions may need to be adjusted over time or adapted to particular country and language contexts.

Overall, we developed the questionnaire with systematic human decision-making in mind. In order to support quick "triage-style" decision-making, a set of initial key questions have been highlighted as potentially being able to support a short-

cut analysis. However, in cases of subtler claims or narratives, the full set of questions may help illustrate whether harmful content is compounding. We will be gathering feedback over the next months from fact-checkers on whether these questions can in fact work in practice.

Depending upon strategic priorities, fact-checking organizations, such as those connected to the International Fact-Checking Network, may be interested in tailoring the questionnaire by focusing on a smaller subset of questions depending on their size and subject focus. In addition, some larger organizations may have access to technologies that can expedite the process for some of these questions.

Determining the threshold for what proportion of "Yes" answers require responses are left up for each individual fact-checking organization to determine.



Urgent Questions

A Structured Response to Misinformation as Harm

Social Futures Lab

In fact checking, a critical need is prioritization. Focusing on the challenge of misinformation, how can fact checkers be supported in their decision making to decide which pieces of content may have more negative impacts in comparison to others? Five major dimensions can help determine the potential urgency of a specific message or post: **actionability**, **exploitativeness**, **likelihood of spread**, **believability**, and **social fragmentation**. A longer version of the questionnaire with helpful hints for answering can be found at: https://artt.cs.washington.edu/online-misinfo-harms-questionnaire/

Urgency Dimension 1: Actionability

A piece of content is more harmful the more that it spurs actions that directly cause harm. Therefore, a piece of misinformation is more harmful the more that it spurs direct action.

Mark a "1" for each answer and tally them up.

Suggestions for hints on how to answer these questions can be found at https://artt.cs.washington.edu/online-misinfo-harms-questionnaire/.

Use the Key Questions if a quick assessment is needed highlighted in (dark blue).

"Don't Know/Not Applicable" (?) can be subtracted from Total "Yes" (Y) answers to gain a relative magnitude of urgency across different pieces of content.

Actionability Questions	Y	N	?
Does the message content include an explicit call to action?			
Does the piece of content incorporate coordination efforts, such as dates/times or other arrangements for follow-up?			
Does the message provide a name or otherwise any identifying information about an individual, an address, or a place of work in such a way that people might be directly harmed?			
Does the message content include a tone of urgency or mention of time sensitivity?			
Does the message content include any threats of violence?			
Does the message lay blame or cast aspersions or hatred on a particular group, such as a particular religion, gender, sexual orientation, race, country, or culture, that has been harmed in the past by the audience of the content?			
Does the message invoke a sense of injustice or moral outrage, including on behalf of a vulnerable individual or group such as children or women?			
Does the direct target or current audience members directly addressed of the message have a recent history of taking actions that cause harm?			
Is this message associated with/similar to other messages that are also actionable?			
Subtotal:			

Urgency Dimension 1: Actionability 01 September 2022

Online Misinformation Harm Questionnaire: A full appendix with all questions in the questionnaire is included at the end of this paper. An online version can be found at: https://artt.cs.washington.edu/online-misinfo-harms-questionnaire/.

Overview: Five Dimensions of Urgency

Dimension 1: Actionability

Questions tied to this dimension are intended to ascertain whether characteristics or factors related to the message or messages make the content likely to spur directly harmful actions. For example, an explicit "call to action" is a key example of actionability, though there are ways that it can be obscured. Additional questions, beyond those that are labeled a 'key,' attempt to capture subtler considerations with regards to actionability.

Overall, the 'actionability' category favors the potential for physical harm over other types, a characteristic recognized both in economic risk assessments and harm evaluations.[22] The focus on physical harm, when considering actionability, was affirmed in our conversations with fact-checkers.

A piece of content that is harmful becomes more harmful when it spurs direct action.

Therefore, a piece of misinformation is more harmful the more that it spurs direct action.

Key questions regarding 'actionability' include:

- Does the message content include an explicit call to action?
- Does the piece of content incorporate coordination efforts, such as dates/times or other arrangements for follow-up?
- Does the message provide a name or otherwise any identifying information about an individual, an address, or a place of work in such a way that people might be directly harmed?

Dimension 2: Exploitativeness

These questions addressing 'exploitativeness' recognize that factors ranging from emotional manipulation to a lack of available resources can contribute to a group's vulnerability to misinformation. Harm frameworks that note the vulnerability of groups such as children are related but focus on characterizing the group, whereas this dimension

[22] Agrafiotis et al., "Cyber Harm"; Scheuerman et al., "A Framework of Severity for Harmful Content Online."

strives to examine when aspects of the message itself directly engage in exploitation.[23]

A piece of misinformation is more harmful the more the message seeks to exploit human or a group's weaknesses, including a lack of resources.

Key questions include:

- Does the message directly address or reference children or use language aimed at a younger audience?
- Does the message directly address or reference elderly community members, or discuss topics aimed at them?
- Does the message introduce a degree of fear or feelings of uneasiness?
- Is the message content complicated to understand?

Dimension 3: Likelihood of Spread

These questions try to ascertain whether characteristics or factors related to the message(s) make the content likely to spread or discoverable. It focuses on questions related to magnitude of exposure or potential exposure rather than analyzing the message for its credibility. Misinformation literature often focuses on this vector when thinking about potential impact and, in our interviews, fact-checkers mentioned virality often when considering their own evaluation of a claim's urgency.[24]

A piece of harmful content is more harmful the more places it appears, and the more people who are exposed to it.

Therefore, a piece of misinformation is more harmful the more places and people are exposed to it.

[23] Scheuerman et al., "A Framework of Severity for Harmful Content Online."

[24] Starbird, Arif, and Wilson, "Disinformation as Collaborative Work."

Key questions include:

- Do the people or entities who are spreading the piece of content have a broad reach (size of following on social media, "influencer," presence on TV or other news media)?
- Are the people or entities known to be repeat spreaders of questionable information?

Dimension 4: Believability

These questions are related to topics where either authoritative consensus is difficult to achieve, or such consensus is affected by the perceptions from a specific community ("in-group"). Answering questions surrounding believability will require at times having a specific community in mind.

A piece of misinformation is more harmful the more believable its message is to a specific community.

Related:

A piece of content is more effective the more believable its message is to a specific community.

Key questions include:

- Is there a lack of high-quality information that is publicly accessible and is refuting the message's claim?
- Does the poster and/or organization/outlet have a noteworthy number of social media/community followers?
- Is the content published by an organization/outlet with uncertain editorial control (e.g., is not a recognized news publisher)?

Dimension 5: Social Fragmentation

This set of questions address the potential of misinformation to affect larger, community-based relationships over time. Issues of peer-to-peer and institutional trust are examples, where one of the long-term consequences of misinformation is reduced trust in existing institutions and social groups. This category emerged out of our exchanges with fact-checkers and our work on a related project on trust.[25] Because

[25] The Analysis and Response Toolkit for Trust (ARTT), at https://artt.cs.washington.edu.

these questions have to do with longer-term implications, there are no shortcut questions.

> *A piece of misinformation could have indirect, societal, and accumulative effects.*
>
> *Therefore, a piece of misinformation is more harmful the more that it undermines societal and community relationships over time.*

Example questions include:

- Does the message fit into a larger narrative that has been existing for some time?
- Does the message question trust in or the functioning of public institutions?
- Does the message question the trustworthiness of other people in general within a community or society?

Misinformation Prioritization: Example Assessment

We share here two examples to demonstrate how the proposed framework could be used in evaluating what content to prioritize in fact-checking.

Take, for example, two content scenarios:

Scenario 1:

A post being shared on multiple platforms that claims that women who take a COVID-19 vaccine cannot get pregnant.

Scenario 2:

A post made by a political candidate accusing a rival candidate standing for political office of sexual assault.

The questionnaire provides specific questions for each of these dimensions that contribute to a score for each of the five dimensions of urgency.

However, to demonstrate the concept of the assessment, we provide high level results of 'low', 'medium', or 'high' urgency for each dimension to give a sense of what a comparison might look like. Fact checkers would evaluate these claims based on their experience and knowledge of the local context.

Summary of Assessment

Claim 1: Pregnancy and COVID-19	**Claim 2: Political candidate, accusations of sexual assault**
High Degree of Harm: Exploitativeness, likelihood of spread (depending on number of reshares), believability	**High Degree of Harm:** Likelihood of spread
Medium Degree of Harm: Likelihood of spread (depending on number of reshares)	**Medium Degree of Harm:** Social Fragmentation, believability
Low Degree of Harm: Social fragmentation, actionability	**Low Degree of Harm:** Actionability, exploitativeness

(The full Misinformation Prioritization example assessment starts on the next page)

Misinformation Prioritization: Example Assessment

Dimension	Claim 1: Pregnancy and COVID-19	Claim 2: Political candidate, accusations of sexual assault
Actionability A piece of misinformation is more harmful the more that it spurs direct action.	• This claim does not call to act against a specific person or a specific event. • This claim can be rated **low** on actionability.	• This claim is indirectly calling to not vote for an individual. It doesn't however ask for direct violence against the candidate. • This claim can be rated **low** on actionability.
Exploitativeness A piece of misinformation is more harmful the more the message seeks to exploit human or a group's weaknesses, including a lack of resources.	• This claim exploits the fear of not being able to conceive, and the guilt of having contributed to an unhealthy pregnancy. Maternal and child health are complicated issues involving both medical and sociological perspectives that aren't always easy to parse. • This claim can be rated as **high** on exploitativeness.	• This claim is invoking the social perception of sexual misconduct as an especially immoral action, which makes a person unfit for office. It doesn't however appeal to a feeling of fear or community specific values. • This claim can be rated as **low** on exploitativeness.
Likelihood of Spread A piece of misinformation is more harmful the more places and people are exposed to it.	• This post is circulated on multiple platforms but is not led by a person of influence, but it might be re-shared widely. • Depending on the number of re-shares this may be rated as **medium** or **high**.	• This post is shared by a person of note with a large online and offline following. • This post can be rated as **high** on likelihood of spread.
Believability A piece of misinformation is more harmful the more believable its message is to a specific community.	• Given the complexity of parsing changing medical evidence, this claim can be difficult to investigate. • This claim can be rated as **high** on believability.	• Allegations of misconduct are hard to prove or disprove by citizens and are in the purview of law enforcement. Since the claim was made by a political candidate about a competitor, the public might be skeptical of the intentions behind the claim. • Thus, the claim can be rated as **medium** believability.
Social Fragmentation Does the content affect societal and community relationships over time?	• This post does not target a specific community or further divide along racial / communal lines. • This claim can be rated as **low** on social fragmentation.	• In a polarized political environment, this post can further the divide between political camps. • This claim can be rated as **medium** social fragmentation.

Notes On the Example Assessment

Claim 1 was ranked as high degree of harm on two dimensions, and medium degree of harm on one dimension. Claim 2 was ranked as high degree of harm on one dimension and medium degree of harm on two

dimensions. With each dimension given equal consideration, Claim 1 emerges as the more harmful claim meriting more urgent attention. It is however possible that fact checkers might weigh one dimension, say social fragmentation, as more important than others. In addition, a continued analysis across time may demonstrate that a set of related claims compound. In this case the evaluation of what needs to be prioritized could change.

Conclusion and Next Steps

This work on misinformation assessment offers a starting point in considering misinformation as a harm in and of itself within larger information exchange. While the task of fact-checking is difficult and filled with nuance and uncertainty, it is an important activity that is not likely to disappear any time soon.

Our effort is one among many to better support fact-checkers by considering the complicated nature of their work. But the end result hopefully goes beyond, by supporting conversations about what we desire for our information environment overall.

To move forward, we invite improvements to this model. We have shared this as a working paper in order to gain feedback; suggestions for refinement and improvement are welcome at socialfutures@cs.uw.edu.

At the same time, we aim to survey fact-checkers for their thoughts, including to determine whether the key questions are appropriate for a short-cut analysis. Additionally, one potential avenue that can further develop this model is to add the level of individual claims, which is a focus of much fact-checking, to post and message-related assessments.

Acknowledgments

We appreciate the input of our fact-checkers and experts over the past year, as well as others who have given us feedback. In particular, we wish to thank Franziska Roesner, Kate Starbird, Aimee Rinehart, as well as members of the Center for an Informed Public for their feedback. We very much thank key contributions of Skyler Hallinan and Alexandra Bornhoft to this paper as well. Thanks too to Nevin Thompson for his support for the styling and production of this paper's questionnaire.

This work has been part of a larger project, the Analysis and Response Toolkit for Trust (ARTT), supported by the National Science Foundation's Convergence Accelerator program under Award No. 49100421C0037; the ARTT team and the authors of this paper wish to thank National Science Foundation's Convergence Accelerator program for its support.

References

Agrafiotis, Ioannis, Maria Bada, Paul Cornish, Sadie Creese, Eva Ignatuschtschenko, Taylor Roberts, and David Upton. "Cyber Harm: Concepts, Taxonomy and Measurement." Saïd Business School Research Papers. University of Oxford, August 2016. http://www.ssrn.com/abstract=2828646.

Agrafiotis, Ioannis, Jason R C Nurse, Michael Goldsmith, Sadie Creese, and David Upton. "A Taxonomy of Cyber-Harms: Defining the Impacts of Cyber-Attacks and Understanding How They Propagate." *Journal of Cybersecurity* 4, no. 1 (October 16, 2018). https://doi.org/10.1093/cybsec/tyy006.

Amazeen, Michelle A. "Revisiting the Epistemology of Fact-Checking." *Critical Review* 27, no. 1 (January 2, 2015): 1–22. https://doi.org/10.1080/08913811.2014.993890.

Barrett, Paul M. "Who Moderates the Social Media Giants? A Call to End Outsourcing." NYU Stern Center for Business and Human Rights, June 2020. https://issuu.com/nyusterncenterforbusinessandhumanri/docs/nyu_content_moderation_report_final_version/1.

Bell, Emily. "The Fact-Check Industry." *Columbia Journalism Review*, 2019. https://www.cjr.org/special_report/fact-check-industry-twitter.php/.

Bok, Sissela. *Lying: Moral Choice in Public and Private Life*. Knopf Doubleday Publishing Group, 2011.

Ceci, Stephen J., and Wendy M. Williams. "[OPINION] The Psychology of Fact-Checking." *Scientific American*, October 25, 2020. https://www.scientificamerican.com/article/the-psychology-of-fact-checking1/.

Sivek, Susan Currie, and Sharon Bloyd-Peshkin. "Where Do Facts Matter?: The Digital Paradox in Magazines' Fact-Checking Practices." *Journalism Practice* 12, no. 4 (April 21, 2018): 400–421. https://doi.org/10.1080/17512786.2017.1307694.

Farkas, Johan, and Jannick Schou. Post-Truth, Fake News and Democracy: Mapping the Politics of Falsehood. Routledge, 2019.

Feinberg, Joel. *The Moral Limits of the Criminal Law Volume 1: Harm to Others*. New York: Oxford University Press, 1987. https://doi.org/10.1093/0195046641.001.0001.

Fischer, Sara. "Fact-Checking Goes Mainstream in Trump Era." *Axios*, October 13, 2020, sec. Economy & Business. https://www.axios.com/2020/10/13/fact-checking-trump-media.

Franklin, Bob, and Brian McNair. *Fake News: Falsehood, Fabrication and Fantasy in Journalism*. London: Routledge, 2017. https://doi.org/10.4324/9781315142036.

FullFact. "Towards a Framework for Information Incidents - Paper 3: Levels of Incidents." FullFact, December 18, 2020. https://fullfact.org/media/uploads/framework_paper_3.pdf.

Graves, Lucas. *Deciding What's True: The Rise of Political Fact-Checking in American Journalism*. New York: Columbia University Press, 2016.

Graves, Lucas, Brendan Nyhan, and Jason Reifler. "Understanding Innovations in Journalistic Practice: A Field Experiment Examining Motivations for Fact-Checking." *Journal of Communication* 66, no. 1 (February 1, 2016): 102–38. https://doi.org/10.1111/jcom.12198.

Iyengar, Shanto, and Douglas S. Massey. "Scientific Communication in a Post-Truth Society." *Proceedings of the National Academy of Sciences* 116, no. 16 (April 16, 2019): 7656–61. https://doi.org/10.1073/pnas.1805868115.

Keyes, Ralph. The Post-Truth Era: Dishonesty and Deception in Contemporary Life. New York: St. Martin's Press, 2004.

Lewandowsky, Stephan, Ullrich K. H. Ecker, and John Cook. "Beyond Misinformation: Understanding and Coping with the 'Post-Truth' Era." *Journal of Applied Research in Memory and Cognition* 6, no. 4 (December 1, 2017): 353–69. https://doi.org/10.1016/j.jarmac.2017.07.008.

Perez, Sarah. "TikTok to Flag and Downrank 'Unsubstantiated' Claims Fact Checkers Can't Verify." *TechCrunch*, February 3, 2021. https://social.techcrunch.com/2021/02/03/tiktok-to-flag-and-downrank-unsubstantiated-claims-fact-checkers-cant-verify/.

Scheuerman, Morgan Klaus, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. "A Framework of Severity for Harmful Content Online." *Proceedings of the ACM on Human-Computer Interaction* 5, no. CSCW2 (October 18, 2021): 368:1-368:33. https://doi.org/10.1145/3479512.

Sehat, Connie Moon, and Farah Lalani. "Advancing Digital Safety: A Framework to Align Global Action." White Paper. World Economic Forum, June 2021. https://www3.weforum.org/docs/WEF_Advancing_Digital_Safety_A_Framework_to_Align_Global_Action_2021.pdf.

Shead, Sam. "TikTok Is Luring Facebook Moderators to Fill New Trust and Safety Hubs." *CNBC*, November 12, 2020, sec. Technology. https://www.cnbc.com/2020/11/12/tiktok-luring-facebook-content-moderators.html.

Siwakoti, Samikshya, Kamya Yadav, Nicola Bariletto, Luca Zanotti, Ulas Erdogdu, and Jacob N. Shapiro. "How COVID Drove the Evolution of Fact-Checking." *Harvard Kennedy School Misinformation Review*, May 6, 2021. https://doi.org/10.37016/mr-2020-69.

Snyder, C. R. "Hope Theory: Rainbows in the Mind." *Psychological Inquiry* 13, no. 4 (October 2002): 249–75. https://doi.org/10.1207/S15327965PLI1304_01.

Starbird, Kate, Ahmer Arif, and Tom Wilson. "Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations." *Proceedings of the ACM on Human-Computer Interaction* 3, no. CSCW (November 7, 2019): 1–26. https://doi.org/10.1145/3359229.

Stencel, Mark, Eric Ryan, and Joel Luther. "Fact-Checkers Extend Their Global Reach with 391 Outlets, but Growth Has Slowed." *Duke Reporters' Lab* (blog), June 17, 2022. https://reporterslab.org/fact-checkers-extend-their-global-reach-with-391-outlets-but-growth-has-slowed/.

Tran, Thi, Rohit Valecha, Paul Rad, and H. Raghav Rao. "Misinformation Harms: A Tale of Two Humanitarian Crises." *IEEE Transactions on Professional Communication* 63, no. 4 (December 2020): 386–99. https://doi.org/10.1109/TPC.2020.3029685.

"Transparency Report 2021 - Reddit." Accessed July 25, 2022. https://www.redditinc.com/policies/transparency-report-2021-2/.

The Hindu. "Twelve Taken Ill after Consuming 'Coronavirus Shaped' Datura Seeds." April 7, 2020, sec. Andhra Pradesh. https://www.thehindu.com/news/national/andhra-pradesh/twelve-taken-ill-after-consuming-coronavirus-shaped-datura-seeds/article31282688.ece.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. "The Spread of True and False News Online." *Science* 359, no. 6380 (March 9, 2018): 1146–51. https://doi.org/10.1126/science.aap9559.

Wardle, Claire, and Hossein Derakhshan. "Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making." Council of Europe, October 2017. https://rm.coe.int/information-disorder-report-november-2017/1680764666.

"YouTube Community Guidelines Enforcement – Google Transparency Report." Accessed July 25, 2022. https://transparencyreport.google.com/youtube-policy/removals?hl=en.

Appendix: Online Misinfoharms Questionnaire

In fact-checking, a critical need is prioritization. Focusing on the challenge of misinformation, how can fact checkers be supported in their decision making to decide which pieces of content may have more negative impacts in comparison to others?

Our model includes five dimensions of urgency when it comes to assessing and prioritizing misinformation as a harm:

1. Actionability
2. Exploitativeness
3. Likelihood of spread
4. Believability
5. Social fragmentation

The five dimensions of urgency are each defined through a set of questions, which are incorporated into a single questionnaire; the questions that make up the questionnaire reflect factors that are currently understood to have an impact upon misinformation's magnitude or potential harm.

The questionnaire is reproduced in the appendix; the PDF version of this questionnaire can be downloaded at this link:

https://artt.cs.washington.edu/misinfoharms-questionnaire/

A longer version of the questionnaire with helpful hints for answering can be found at:

https://artt.cs.washington.edu/online-misinfo-harms-questionnaire/

Urgent Questions

A Structured Response to Misinformation as Harm

Analysis and Response Toolkit for Trust

Social Futures Lab

In fact checking, a critical need is prioritization. Focusing on the challenge of misinformation, how can fact checkers be supported in their decision making to decide which pieces of content may have more negative impacts in comparison to others? Five major dimensions can help determine the potential urgency of a specific message or post: **actionability**, **exploitativeness**, **likelihood of spread**, **believability**, and **social fragmentation.** A longer version of the questionnaire with helpful hints for answering can be found at: https://artt.cs.washington.edu/online-misinfo-harms-questionnaire/

Urgency Dimension 1: Actionability

A piece of content is more harmful the more that it spurs actions that directly cause harm. Therefore, a piece of misinformation is more harmful the more that it spurs direct action.

Mark a "1" for each answer and tally them up.

Suggestions for hints on how to answer these questions can be found at https://artt.cs.washington.edu/online-misinfo-harms-questionnaire/.

Use the Key Questions if a quick assessment is needed highlighted in (dark blue).

"Don't Know/Not Applicable" (?) can be subtracted from Total "Yes" (Y) answers to gain a relative magnitude of urgency across different pieces of content.

Actionability Questions	Y	N	?
Does the message content include an explicit call to action?			
Does the piece of content incorporate coordination efforts, such as dates/times or other arrangements for follow-up?			
Does the message provide a name or otherwise any identifying information about an individual, an address, or a place of work in such a way that people might be directly harmed?			
Does the message content include a tone of urgency or mention of time sensitivity?			
Does the message content include any threats of violence?			
Does the message lay blame or cast aspersions or hatred on a particular group, such as a particular religion, gender, sexual orientation, race, country, or culture, that has been harmed in the past by the audience of the content?			
Does the message invoke a sense of injustice or moral outrage, including on behalf of a vulnerable individual or group such as children or women?			
Does the direct target or current audience members directly addressed of the message have a recent history of taking actions that cause harm?			
Is this message associated with/similar to other messages that are also actionable?			
Subtotal:			

Urgency Dimension 2: Exploitativeness

A piece of misinformation is more harmful the more the message seeks to exploit human or a group's weaknesses, including a lack of resources.

Mark a "1" for each answer and tally them up.

Suggestions for hints on how to answer these questions can be found at https://artt.cs.washington.edu/online-misinfo-harms-questionnaire/.

Use the Key Questions if a quick assessment is needed highlighted in (dark blue).

"Don't Know/Not Applicable" (?) can be subtracted from Total "Yes" (Y) answers to gain a relative magnitude of urgency across different pieces of content.

Exploitativeness Questions	Y	N	?
Does the message directly address or reference children or use language aimed at a younger audience?			
Does the message directly address or reference elderly community members, or discuss topics aimed at them?			
Does the message introduce a degree of fear or feelings of uneasiness?			
Is the message content complicated to understand?			
Does the message directly address or reference military veterans, or discuss topics aimed at them?			
Does the message make mention of a reader's feelings of isolation?			
Does the message make mention of a reader's feelings of powerlessness?			
Does the message make mention of a reader's feelings of disenfranchisement?			
Is this message associated with/similar to other messages that are also actionable?			
Is the language of the intended audience neither a UN language (English, French, Spanish, Mandarin Chinese, Russian) nor on the top 5 list of most popular languages?			
Is the message presented in a region where the local context might amplify its harm?			
Subtotal:			

Urgency Dimension 3a: Likelihood of Spread

A piece of harmful content is more harmful the more places it appears, and the more people are exposed to it. Therefore, a piece of misinformation is more harmful the more places and people are exposed to it.

Mark a "1" for each answer and tally them up. Suggestions for hints on how to answer these questions can be found at https://artt.cs.washington.edu/online-misinfo-harms-questionnaire/.

Use the Key Questions if a quick assessment is needed highlighted in (dark blue).

	Likelihood of Spread Questions	Y	N	?
WHO is spreading?	Is the content already spreading far and/or fast on a multitude of platforms?			
	Do the people or entities who are spreading the piece of content have a broad reach (size of following on social media, "influencer," presence on TV or other news media)?			
	Are the people or entities known to be repeat spreaders of questionable information?			
WHERE is it spreading?	Is there evidence of coordination activity (whether bot/automated or not) to encourage spread?			
	Is the content publicly accessible (posted on a public platform, addressable URL)? Is the content posted on a popular platform?			
	Is the content spreading on multiple platforms?			
	Does one of the platforms upon which the content is shared have tools to support amplification (e.g. reshares, algorithmic feeds, recommendation engines)?			

...Urgency Dimension questions continue on next page

"Don't Know/Not Applicable" (?) can be subtracted from Total "Yes" (Y) answers to gain a relative magnitude of urgency across different pieces of content.

Urgency Dimension 3b: Likelihood of Spread

A piece of harmful content is more harmful the more places it appears, and the people are exposed to it. Therefore, a piece of misinformation is more harmful the more places and people are exposed to it.

Mark a "1" for each answer and tally them up. Suggestions for hints on how to answer these questions can be found at https://artt.cs.washington.edu/online-misinfo-harms-questionnaire/.

Use the Key Questions if a quick assessment is needed highlighted in (dark blue).

"Don't Know/Not Applicable" (?) can be subtracted from Total "Yes" (Y) answers to gain a relative magnitude of urgency across different pieces of content.

CHARACTERISTICS of the message

Likelihood of Spread Questions	Y	N	?
Does the message make direct appeals to audience members that it in their financial, political, or social interest to spread the content further?			
Does the message directly call audience members to share the content further?			
Is the tone of the content striking enough in ways that encourage sharing?			
Does the content contain an image, audio-clip, or other richer formats that are easy to remember, visually or aurally arresting, or seems interesting to share?			
Does the message impart a sense of exclusivity or novelty (“breaking news”)?			
Are there hashtags associated with the message?			
Is the message difficult to fact-check or prove false?			
Is the message related to a current event or a topic that is being reported on actively by many news outlets ?			
Subtotal:			

Instructions:
Please tally 3a and 3b totals in Subtotal, above right.

Urgency Dimension 4: Believability

A piece of misinformation is more harmful the more believable its message is to a specific community.

Mark a "1" for each answer and tally them up.

Suggestions for hints on how to answer these questions can be found at https://artt.cs.washington.edu/online-misinfo-harms-questionnaire/.

Use the Key Questions if a quick assessment is needed highlighted in (dark blue).

"Don't Know/Not Applicable" (?) can be subtracted from Total "Yes" (Y) answers to gain a relative magnitude of urgency across different pieces of content.

Believability Questions	Y	N	?
Is there a lack of high quality information that is publicly accessible and is refuting the message's claim?			
Does the poster and/or organization/outlet have a noteworthy number of social media/community followers?			
Is the content published by an organization/outlet with uncertain editorial control (e.g. is not a recognized news publisher)?			
Does the message fail to include external citations, links, or language about evidence to support its claim?			
Does the message contain richer formats as part of its evidence that lay people consider to have low falsifiability?			
Is the message written or communicated in a personal or persuasive tone?			
Does the message make reference to the broad believability of the claim or topic?			
Does the message appeal to a specific community identity by mentioning a shared set of values or beliefs?			
Is there a lack of consensus on the part of experts regarding the claim?			
Does the poster have credentials that represents some kind of expertise?			
Is the content posted by an imposter individual or counterfeit outlet that could successfully pass as a different person/account based only upon a quick glance?			
Does the content have the graphics and styling of a legitimate news agency or mainstream information source?			
Subtotal			

Urgency Dimension 5: Social Fragmentation

A piece of misinformation could have indirect, societal, and accumulative effects. Therefore, a piece of misinformation is potentially more harmful the more that it addresses or is part of societal and community relationships over time.

There are no short cut questions for this category.

Mark a "1" for each answer and tally them up.

Suggestions for hints on how to answer these questions can be found at https://artt.cs.washington.edu/online-misinfo-harms-questionnaire/.

Use the Key Questions if a quick assessment is needed highlighted in (dark blue).

"Don't Know/Not Applicable" (?) can be subtracted from Total "Yes" (Y) answers to gain a relative magnitude of urgency across different pieces of content.

Social Fragmentation Questions	Y	N	?
Does the message fit into a larger narrative that has been existing for some time?			
Does the message question trust in or the functioning of public institutions?			
Does the message question trust in or the functioning of the scientific community as a whole?			
Does the message question the functioning of or trust in news sources/ the media in general?			
Does the message question the trustworthiness of other people in general within a community or society?			
In a democratic country where there are elections, does the message directly attack the election process?			
Subtotal			

Urgency Dimension 1: Actionability	Urgency Dimension 2: Exploitativeness	Urgency Dimension 3: Likelihood of Spread	Urgency Dimension 4: Believability	Urgency Dimension 5: Social Fragmentation	TOTAL