Misinformation as a harm: structured approaches for fact-checking prioritization

CONNIE MOON SEHAT, Hacks/Hackers, USA
RYAN LI, Stanford University, USA
PEIPEI NIE, University of Washington, USA
TARUNIMA PRABHAKAR, Tattle Civic Technologies, India
AMY X. ZHANG, University of Washington, USA

In this work, we examined how fact-checkers prioritize which claims to inspect for further investigation and publishing, and what tools may assist them in their efforts. Specifically, through a series of interviews with 23 professional fact-checkers from around the world, we validated that harm assessment is a central component of how fact-checkers triage their work. First, we clarify what aspects of misinformation they considered to create urgency or importance. These often revolved around the potential for the claim to harm others. We also clarify the processes behind collective fact-checking decisions and gather suggestions for tools that could help with these processes.

In addition, to address the needs articulated by these fact-checkers and others, we present a five-dimension framework of questions to help fact-checkers negotiate the priority of claims. Our FABLE Framework of Misinformation Harms incorporates five dimensions of magnitude—*(social) Fragmentation*, *Actionability*, *Believability*, *Likelihood of spread*, and *Exploitativeness*—that can help determine the potential urgency of a specific message or post when considering misinformation as harm. This effort was further validated by additional interviews with expert fact-checkers. The result is a questionnaire, a practical and conceptual tool to support fact-checkers and other content moderators as they make strategic decisions to prioritize their efforts.

CCS Concepts: • **Human-centered computing** → **Collaborative filtering**.

Additional Key Words and Phrases: fact-checking, harm, misinformation, taxonomy, decision-making, virality

ACM Reference Format:
Connie Moon Sehat, Ryan Li, Peipei Nie, Tarunima Prabhakar, and Amy X. Zhang. 2024. Misinformation as a harm: structured approaches for fact-checking prioritization. *Proc. ACM Hum.-Comput. Interact.* N, CSCWN, Article X (2024), 47 pages. https://doi.org/XXXXXXX.XXXXXX

1 INTRODUCTION

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 [10, 30, 38, 42]. 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

Authors' addresses: Connie Moon Sehat, Hacks/Hackers, USA; Ryan Li, Stanford University, USA; Peipei Nie, University of Washington, USA; Tarunima Prabhakar, Tattle Civic Technologies, India; Amy X. Zhang, University of Washington, USA.

2573-0142/2024/00-ARTX $15.00
https://doi.org/XXXXXXX.XXXXXX

political activities such as voting, interfere with the peaceful transition of power, and even motivate political violence [37, 39].

Fact-checking is a critical strategy 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 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. Given that prioritization is unavoidable, how should fact-checking efforts to combat misinformation prioritize what content to tackle? Can the prioritization be systematized? Can a systematic process also reflect the priorities and desires of fact-checkers?

One way forward is through *harm assessment*. Taking the approach that misinformation could be treated as a harm opens up a fruitful line of inquiry, as 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 and helping the public to achieve informed decision-making [22]. In addition, while all misinformation is harmful to some degree, not all misinformation is equally harmful, making harm assessment a potentially useful component of prioritization.

Through a series of interviews with 23 professional fact-checkers from around the world, we validated that harm assessment is a central component of how fact-checkers triage their work. We gained an understanding of how fact-checkers determine harm, including what they look for, how and when they incorporate harm assessment into their process, and the other factors considered, when prioritizing what to fact-check. We also wanted to understand the role of tools, existing and proposed, that could support this process. In summary, we sought the answers to the following research questions:

- RQ1: According to fact-checkers, what aspects of misinformation create urgency or importance?
- RQ2: How do fact-checkers decide what to fact-check and what tools could improve their processes of prioritization?

From our interviews, we discover that fact-checkers take many considerations into account when they prioritize what claims should be fact-checked. Key among their concerns is the potential harmfulness of a claim (particularly when it is physical), the claim's likelihood of spread or virality, and the potential impact of a fact-check. We also find that fact-checking processes, overall, are still young and not standardized. Fact-checkers typically take a relatively ad hoc approach to prioritization, using individual judgment and case-by-base discussion with others. Regarding tools, fact-checkers desire features that can help ease their work or speed up their processes, as well as tools that help them assess the potential harmful impact of misinformation in ways that are sensitive to local context.

Drawing on these findings, we present a novel *misinformation harms framework* to enable fact-checkers with prioritization in a more structured fashion. Following a literature review, workshops with fact-checkers and other misinformation experts, and an incorporation of the interview findings, we developed dimensions of analysis to help prioritize fact-checking efforts within the format of a questionnaire. Using a draft of the taxonomy and accompanying questionnaire, we received feedback from 4 additional professional fact-checkers, and iterated again on the dimensions and questions.

Our FABLE Framework of Misinformation Harms incorporates five dimensions of magnitude—*(social) Fragmentation*, *Actionability*, *Believability*, *Likelihood of spread*, and *Exploitativeness*—that can help determine the potential urgency of a specific message or post when considering misinformation as harm. The framework, and its questions, are intended as both conceptual and practical tools that, based on the desires and perspectives of fact-checkers, may support them, content moderators, peer correction efforts, and other initiatives as they make strategic decisions when prioritizing their efforts to respond to misinformation that is spreading. We discuss ways our framework and questionnaire could be used within misinformation response practice and also discuss design implications for tools to support misinformation response.

2 RELATED WORK

2.1 Harms of Misinformation

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. There are moral dilemmas behind lying at an individual and social level, even for just the "harmless white lie" [9]. For the last two decades, concerns from journalists, political scientists, cognitive psychologists, and other research communities about whether we are in a "post-truth" society 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 [15, 19, 25, 28, 29, 46].

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 [43]. In online realms a wide range of socially undesirable content, such as harassment, child exploitation and narratives leading to self-harm, can be characterized as harmful. 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[12, 16].

Practically speaking, existing practices for addressing potentially harmful content face the challenge of triage. In particular, content moderators and fact-checkers must contend with a tidal wave of content shared 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 [7]. Reports include people working to review between 25 to 100 pieces of content every hour [7, 44]. More examples of this scale and challenge can be seen in reports from other platforms such as YouTube or Reddit [3, 4].

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.[1] For a variety of reasons, it may be more urgent to try to combat the latter example of misinformation.

2.2 Fact-checking Practices

Online platforms that focus on user-generated content have been increasingly exploring ways to scale content review to support safe and accountable exchanges of information, in order to match the pace of online distribution. This is one reason for an explosion of growth in the field of

[1]See for example https://en.wikipedia.org/wiki/Death_hoax in contrast to [2]

fact-checking, which originally grew out of magazine journalism in the earlier half of the twentieth century [21]. Fact-checking has emerged as a distinct profession, independent of journalistic training or traditional journalism channels such as newspapers—for example, a recent census counted 391 fact-checking organizations in 2022 in contrast to 186 in 2016 [47]. In addition, the demand for accurate information in elections and health contexts itself has changed the nature of fact-checking work [17, 45].

To understand how well fact-checking organizations have adapted to these demands, researchers have unpacked the fact-checking process by revealing the human and technological infrastructures that support and shape fact-checking work [27], and surfaced a pipeline of practices fragmented across disparate tools that lack integration [32]. Juneja and Mitra find that fact-checking is more than one-off debunking of misleading claims, but also involves long-term advocacy work to improve the information ecosystem [27]. This finding resonates with Micallef et al.'s work, where their participants mentioned that they contribute to the information ecosystem to facilitate the creation of a balanced public sphere for discussing issues [32]. Also worthy of consideration are the developments in less professionalized contexts. Work in collaborative and crowdsourced approaches to fact-checking reveal similar concerns and discussions, where the processes of this voluntary work again goes well beyond simple debunking [24, 54].

With regard to increasing demands for scaling fact-checking practices through automated tools, researchers find that the largely manual and labor intensive nature of current fact-checking practices is a barrier to scale [32]. This resonates with other work, which argues that quality data is essential not only for developing AI-based automated tools but also for investigating claims [27]. Given the difficulty of scaling up fact-checking, our work considers the question of triage, so that careful human efforts may go towards those areas that may have the greatest impact. In our work, we shed additional light into the practices, struggles, and needs of fact-checkers by interviewing a diverse set of fact-checkers about how they consider prioritization of what to fact-check in their work.

2.3 Structured Harm and Misinformation Assessments

There are a number of works that have attempted to assess harm or misinformation in structured ways. Examples can be found in cybersecurity literature thinking about harms, such as economic, social, and even reputational harms. Other frameworks coming from a content moderation perspective focus on either misinformation or harm, with the weight of harm- versus misinformation-related definitions depending on the purpose of the classification. Here we present an overview of many of the taxonomies in this space, while in Section 4, we dive into a few of the most relevant frameworks in more detail to discuss how their specific dimensions relate to the dimensions we developed in our misinformation harms framework.

2.3.1 Misinformation-oriented taxonomies. We first describe taxonomies of misinformation, most of which focus on categorizing and characterizing different forms of misinformation according to their content or goal. Fitzgerald et al. in 1997 was probably one of the earliest approaches to evaluate online misinformation, identifying 10 types of online misinformation [18]. A more recent taxonomy identified by the First Draft organization, with authors Wardle and Derakshan, has become widely adopted [51–53]. The main focus of their conceptual definitions was to characterize false and misleading information, though their additional category of 'malinformation', or factual information used to inflict harm, does intersect with harms-related considerations. Nakamura et al. followed this work to present a dataset of Reddit posts classified under Wardle's seven types of fake news [36]. Other attempts include McCright et al., which proposes four high level types—truthiness, bullshit, systemic lies, and shock-and-chaos [31].

Researchers have used a range of methods to characterize misinformation using other lenses to develop alternative taxonomies. Jiang et al. focused on misinformation stories and came up with 10 categories by studying the archived fact-checks from Snopes.com [26]. Charquero-Ballester et al. categorized COVID-19 misinformation into six types, through a focus on their claims or narratives, and compared the emotional valence among the different types [14]. In contrast, Brennen et al. focused on identifying the common formats, sources, and claims of COVID-19 misinformation [11]. Psychological studies of misinformation have also differentiated between neutral versus non-neutral misinformation [35], and contradictory versus additive misinformation [34].

2.3.2 Harm-oriented taxonomies. Separately, there are also many taxonomies focused on online harms broadly construed. Agrafiotis et al. have defined a taxonomy of harms from a cybersecurity perspective, where dimensions of harm defined include economic, social, and reputational harms [5, 6]. Recent works have also studied the types and targets of offensive online content or online hate for purposes including content moderation. Zampieri et al. presented a three-layer annotation scheme for detecting offensive online contents [55]. The paper labels offensive content by whether it is a targeted insult or untargeted profanity, and whether it is targeting an individual, a group, or something else (e.g., an organization, a situation, an event, or an issue). Thomas et al. presented a taxonomy of seven categories of online threats and attacks (toxic content, content leakage, overloading, false reporting, impersonation, surveillance, & lockout and control) [49]. And Salminen et al. created a granular taxonomy for hateful online comments, describing 4 types of offensive language, 9 types of targets, and 16 types of sub-targets [40]. Finally, most major platforms publish community guidelines that outline specific categories of objectionable content, and some have released more detailed annotation guidelines that they provide to their paid content moderation staff.

2.3.3 Urgency- or severity-related taxonomies of misinformation harms. Finally, most relevant to our work are taxonomies that bring together an evaluation of harms with the context of misinformation, and that have dimensions that can speak to a degree of urgency or severity. First, Scheuerman et al. 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; however the consideration of misinformation only appears within the single category of Coordinating Scams and Political Attacks [41]. More closely related is non-profit organization FullFact's white papers in recent years, which have attempted to establish a framework of severity around 'information incidents,' or large-scale public incidents where the coordination across organizations and institutions. 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 [20]. Furthermore, Tran et al. 2020 investigated the likelihood of occurrence and level of impact of 15 different types of harms related to misinformation, though they were specifically focused on four different scenarios related to humanitarian crises [50]. Finally, though not explicitly incorporating severity, Mirza et al. has recently proposed a cybersecurity-inspired framework that characterizes disinformation threats by four dimensions: threat actors, attack patterns, attack channels, and target audience [33].

As can be seen, while there are many frameworks relating to either misinformation or harms, there are few that focus specifically upon the harms resulting from misinformation and the potential levels of severity or urgency that may arise. In addition to addressing this gap, our work goes a step further to provide practically applicable worksheets that can be used by fact-checkers to systematize their own processes around prioritization.

3 INTERVIEW STUDY

When embarking on research around misinformation as a harm, we found few public resources or research discussing how fact-checkers conduct prioritization. In order to validate our own perception of harm as a main factor in assessing the urgency of addressing misinformation, we decided to gather direct experiences and perceptions of fact-checkers. We conducted semi-structured interviews with professional fact-checkers from certified fact-checking organizations that explored various aspects of their work—not only how much notions of potential harm motivated their own processes but also their processes of triage. We focus on professional fact-checkers compared to volunteers since the contractual nature of their work makes the question of triage more pressing. In addition, we wanted to know more about how their work might be made easier with new or better tools.

3.1 Method

This section describes participant recruitment, our study protocol, characteristics of our participant sample, and how we conducted analysis (Figure 1). All study procedures were approved by a university's Institutional Review Board (IRB).

3.1.1 Participant Recruitment. We recruited a total of 23 participants who have experience with fact-checking and are working within a fact-checking organization or team. 12 participants were recruited from personal connections of our collaborators who were working for news agencies. 11 participants were recruited through cold emailing—in September 2021, we sent 53 emails to fact-checking organizations certified by the International Fact-Checking Network (IFCN). We manually collected these email addresses from the IFCN website. In our email, we provided a web page describing the study, with the link at the end for people to sign up. We selected participants to interview from our sign-up form in order to have representation from a diversity of fact-checking organizations working in different countries around the world. We then emailed participants and scheduled a time to meet with them over Zoom.

3.1.2 Study Protocol. We interviewed participants individually using a semi-structured protocol that covered the following themes: The background information about the fact-checkers and their organizations, their typical fact-checking process, how they make decisions about prioritization, both as individuals and as part of an organization. We also asked how they considered harms when prioritizing, and what aspects of misinformation they look for when assessing harms. Finally, we asked about their use of or interest in computational or other tools to support prioritization. All interviews took place over the conferencing software Zoom. Interviews lasted around 60 minutes each and were conducted in English. All participants were compensated with a $35 gift card; however, four participants declined the compensation.

3.1.3 Participant Demographics. Table 1 provides an overview of our participant sample. The 23 participants (8 women and 15 men) were from 15 countries covering Africa, Asia, Europe, Latin America, North America, and Oceania. The sample included fact-checkers from a mix of organizations, with 1 from small fact-checking organizations (3-6 employees), 4 from medium fact-checking organizations (7-12 employees), and 18 large fact-checking organizations (more than 12 employees). We omit the names of the fact-checking organizations in order to preserve the anonymity of our participants.

3.1.4 Data collection and analysis. We recorded all interviews with participant permission and transcribed them for analysis. We analyzed the transcripts using a Grounded Theory approach [48]. This approach allowed common themes to emerge from the data in an inductive and interpretative

#	Gender	Country	Role
P1	Male	South Africa	Research role
P2	Female	Ukraine	Editing role
P3	Male	India	Organizational lead
P4	Male	India	Organizational lead
P5	Female	South Africa	Fact-checker
P6	Male	Greece	Editing role
P7	Male	France	Fact-checker
P8	Male	India	Fact-checking lead
P9	Female	Australia	Editor
P10	Female	Argentina	Organizational lead
P11	Male	United States	Fact-checker
P12	Male	Taiwan	Organizational lead
P13	Male	Italy	Organizational lead
P14	Female	Sweden	Organizational lead
P15	Female	France	Trainer
P16	Female	Ukraine	Editing role
P17	Male	Poland	Organizational lead
P18	Male	Brazil	Organizational role
P19	Male	Nigeria	Editing role
P20	Male	France	Editing role
P21	Male	France	Organizational lead
P22	Male	Spain	Coordinator
P23	Female	India	Organizational role

Table 1. Demographic details of interviewees for this study.

manner. Specifically, we randomly selected three transcripts, and two of our authors then open coded them independently. During the open coding phase, the two authors coded the data on a sentence-by-sentence basis and created codes without initial hypotheses. They regularly came together to discuss and resolve disagreements on the codes. Subsequently, they examined the codes for similarities, removed the redundant codes, and created a codebook with definitions for each code. All the authors also reviewed the codebook and discussed the definitions and possible overlapping codes. Then, the two authors went on to split up the remaining transcripts and independently coded them, while continuing to discuss and iterate on the shared codebook with each other and the full team as new codes arose. We reached theoretical saturation after analyzing 18 out of the 23 interviews as no new codes emerged after that point. See Appendices B and C for the full results. These thematic codes resulted in the answers to our RQs, discussed below.

3.2 RQ1: According to fact-checkers, what aspects of misinformation create urgency or importance?

We begin with a deep focus upon how fact-checkers prioritize their review of inaccurate and contested claims, leaving aside for now questions related to their processes. Specifically, we dive into the aspects that many of our interviewees mentioned when they gauge the importance and urgency around misinformation. We also describe how fact-checkers talked about misinformation playing out differently in different regions of the world.

3.2.1 Whether the misinformation may lead to different types of harmful impact. Interviewees often reflected upon or differentiated between two following types of harmful impact when considering

urgency of misinformation. If a piece of misinformation can be argued to possibly cause negative physical or societal effects, which can at times be intertwined, then interviewees considered it important to address.

- **Physical harm.** All participants reported that they would prioritize fact-checking misinformation that has potential for physical harm and consider it the biggest threat of misinformation. One form of physical harm that was repeatedly mentioned was the use of misinformation to provoke violent attacks by stoking outrage or calling for retribution against a group. The other kind of physical harm mentioned was misinformation that may lead people to make poor health choices. Participants highlighted vaccine misinformation in particular as urgent because of its link to physical harm, as misleading information may convince people against getting a possibly life-saving vaccine:
  "*Oftentimes when things are circulating a lot online, it spills into real life. The thing about a vaccine is also urgent for the same reason, because the more you read about the problem with the vaccine online, then you just don't want to get it.*" (P9)
  Some participants also justified the urgency of addressing certain misinformation using the reasoning of physical harm even if the effect was not necessarily direct or immediate. For instance, misinformation that casts doubt on climate change could cause significant physical harm as a second-order effect, as disbelief about climate change could lead to inaction on policy, which then leads to more climate-related deaths.

- **Societal harm.** Another form of harm that was highly referenced was societal harm, or harm that adversely impacts the cohesion and functioning of a society. For instance, a particularly prominent kind of misinformation that our participants considered urgent was election misinformation, which does not necessarily directly cause physical harm. However, similar to the case of climate misinformation, some participants still linked societal harm to physical harm as a second-order effect:
  "*It doesn't really kill you even if you believe Donald Trump has won, but it threatens democracy, which is the pillar of our society. Then you see people were organizing to go to the Capitol before January 6, right? They were talking about that on...all those forums. And then, there were people who died in the thing.*" (P9)

As can be seen, physical and societal harm can be deeply intertwined, and in many cases, interviewees felt that one will also imply the other. However, of the two, interviewees signaled the greater importance of physical harm. Not only did all our participants mention potential physical harm of misinformation as a priority, but individual interviewees highlighted its direct and immediate impacts as the underlying rationale for urgent address.

3.2.2 Characteristics of the misinformation content. In addition to referencing whether the possible impacts of a piece of misinformation, our interviewees talked about specific ways that the misinformation itself was presented or framed that might make it more or less important to address. Most of the time, they described characteristics of the content that would make it more likely for the audience to believe and then act on the content, thus potentially leading to physical or societal harm, or share the content further, thus exposing more people to the misinformation.

- **Emotional and sensational language.** Almost all participants mentioned the role that emotional language and sensationalism play in the virality of misinformation. People writing misinformation content will often use tactics to stoke emotions such as outrage with the goal of getting it shared further:

> "*I guess if it appeals to the emotions in a certain way, this is something that I see frequently, where the poster will say whatever claim, and 'This should make you angry,' and 'I can't believe this is happening,' things along those lines. Just framing misinformation in such a way that appeals to emotions I think makes it much more likely that this piece of misinformation will then get reshared.*" (P11)

Another tactic is to use language that sensationalizing in order to convey a sense of urgency or danger to the reader, which will encourage them to share the content further or act on the misinformation in some way:

> "*Look, people do tend to respond to claims that are more sensationalized...And claims that use capital letters more to make themselves sound more urgent...the claims that tend to go viral are the ones that tend to be very sensationalized like 'Oh my God guys, we are going to die.'*" (P5)

Interviewees pointed out that this tactic isn't always indicative of an urgency to address on its own. Indeed, P3 pointed out that if it is too obvious that "*someone is trying to instigate something by pitting people against each other*," people might catch on. This is similar to the above quote from P2 about anti-vaccine misinformation, where too much use of sensationalism that doesn't eventually pan out may lead to loss of credibility.

- **Time sensitivity.** Most of our participants mentioned instances when a piece of misinformation is conveying a sense of time sensitivity. This was considered important to address because the ways that such messages encourage panic and, thus, virality. One interviewee talked about cases where the speaker urges the audience to perform certain actions that are claimed to have crucial stakes within a short time-frame:

  > "*For instance, this store is closing down or is going bankrupt or being looted or something like that, you need to stock up on this thing. That's very time sensitive because it gives people a time limit and it tells them to go and do something and take some action.*" (P1)

3.2.3 The information context in which the misinformation is interpreted or spread by the public. Interviewees also talked about how urgency can be higher or lower based on its context, or how the piece fits into the landscape of other content and claims. When a piece of misinformation taps into a narrative such that people see connections with past happenings, or is responsive to current events and ongoing discussions, it can have outsized impact due to greater receptiveness and contagion within the public—as more people are exposed to the claim as it travels, the more potential for harm. In addition, misinformation can fill a void where there is currently a lack of public information or significant confusion.

- **Part of a larger narrative or body of misinformation.** While millions of fake or misleading claims circulate around the internet every day, many of our participants pointed out that many individual messages fit into a larger narrative. And together, the larger narrative imposes a more profound impact on our society than any of the individual messages—in other words, misinformation has accumulative effects. Thus, even if an individual claim itself does not directly or immediately lead to physical or societal harm, it may be more important to address if it contributes to a narrative with physically or socially harmful consequences. For instance, when a piece of content is put into a larger narrative such as about anti-vaccination, even an apparently innocent joke could become harmful:

  > "*When you may fit this joke into the bigger narrative ... for example, there is a picture of Mike Tyson wearing a black t-shirt, and it has the emblem [of] something related to anti-vaccination. On one side, it's just a doctored picture, but on the other side, you may fit it*

into the bigger narrative that, look, Mike Tyson, he supports the anti-vaccination movement. Then anti-vaccination ideas get reinforced." (P2)

Some interviewees also talked about even broader and more longer-term narratives such as age-old biases or prejudices about certain faiths or communities. In addition, if there is still existing tension between different groups, such between the different castes and religions within India, misinformation can be considered more urgent to address:

"*...especially there have been riots in the country in the past between people from different faiths...when such [a] thing happens, it can again increase ongoing real-world violence...a very simple child kidnapping rumor in India has caused multiple deaths.*" (P8)

Most interviewees also brought up the accumulative effects of misinformation as a whole, as the flood of false stories over time can overwhelm our capacity to reason and believe. Over time, different unrelated narratives may overlap as people make connections among them; one example is the QAnon conspiracy that incorporates elements of many different conspiracy theories and racist tropes [8, 57]. Context, therefore, should have some utility for a fact-checker's prioritization criteria, as misinformation that is widely believed and durable over time may arguably be important to address in spite of its lack of direct or immediate impact; some interviewees expressed concerns about the overwhelming nature of competing information in ways that align with research on social media fatigue [56].

- **Relevance to current events and people.** As all but one of our interviewees pointed out, misinformation is seasonal in nature. Like news, misinformation (around a certain topic) becomes relevant when something happens to catch people's attention, and fades out when the public shifts its attention to something else. Thus, beyond the pressure to produce fact-checks that drive traffic to their site, fact-checkers find it is important to fact-check misinformation that ties into current events because people are more likely to share it:

  "*...if you talk today about the COVID-19, it makes sense. But if you talk about HIV or polio or measles, it would not be going as viral because people will not be able to connect with [the] majority of the people. Now COVID-19 is something everyone can connect with.*" (P8)

  Typically, when major news breaks, misinformation related to it also becomes viral as readers strive to make sense of the situation; attention to the false stories eventually fades away as the news topic itself recedes from the center of public attention. Part of the reasoning behind the boom-bust cycle is the desire for novelty both in the news industry and on the part of people:

  "*Now it's about health misinformation. Maybe in a year, if it's all over, it's going to be about something else. I'm really interested in what is going to replace this health misinformation...Now it's an abundance of health misinformation but it's going to fade out eventually. Everyone is bored by all of this. Not a lot of people are damaged by the vaccine. The fakes about 'you are going to die' lose their credibility.*" (P2)

  Helping to assist these trends is when the content is shared by someone who is highly influential or trusted; for instance P9 had this to say about influencers on Instagram and TikTok: "*They can be so powerful, especially on younger people because they grow up using them.*"

- **Lack of accessible public information on a topic.** Interviewees also talked about how sometimes it would be important to address something because of a lack of public information about that topic in the internet or social media landscape, a kind of "information gap." When there's lack of information, for instance, about an emerging topic, it can be a more urgent

situation both because people may be more willing to share it to "get the word out" and also because people may be more vulnerable to believing the misinformation:

> "*...it's playing on confusion or a lack of information somewhere. So for instance, when there is a new strain of COVID 19 discovered, then the misinformation targets that new strain, because people don't know a lot about it yet...So you might see a message that an expert would know is nonsense just by looking at it. But if it's about a topic that the average person won't know much about, then they're going to share it just in case.*" (P1)

In addition to a lack of information, there's also the issue of topics that are just inaccessible for laypeople, such as topics related to science. In this case, there is public information but it requires some interpretation by someone with expertise to properly understand:

> "*Just one of the most common reasons pieces of misinformation have gone viral that I've seen is people who don't have the level of expertise necessary to interpret pieces of information misinterpreting that, and misapplying their understanding of that.*" (P11)

However, not every interviewee found this situation to be a criteria for importance. We noticed that only fact-checkers from regions where science- and health-related misinformation was prevalent ranked the information gap highly, whereas people from places where political, religious and social conflicts were the main issues deemed it as unimportant. The offline or local contexts, in other words, mattered.

3.2.4 Affinity of misinformation for certain communities, cultures, and countries. In fact, the consideration of offline context raised a final theme by interviewees: the recognition that misinformation has different impacts in different communities and areas with different norms and beliefs. There is, in other words, an affinity of certain kinds of misinformation for specific contexts, where such stories are more believable or more likely to go viral, making them more important to address.

For instance, health related misinformation is prevalent in the U.S., and vaccine hesitancy has been a long-standing problem. However, that isn't the case for India, where vaccine hesitancy is not as common. One participant (P4) mentioned that in India, there is a shortage of vaccine supply compared to the demand, and most media convey that people want to get vaccinated. In fact, for countries like India and South Africa, misinformation and hate speech relating to social and religious discrimination has been a much more prominent issue. Understanding the affinity of misinformation requires a clear understanding of the local and cultural contexts on the part of fact-checkers to know which misinformation will have disparate impacts at a local level.

- **Misinformation that attacks marginalized groups.** Almost all interviewees spoke about how misinformation can be deployed to attack specific communities where there are existing local prejudices. Some highlighted misinformation targeting specific minority religious groups or marginalized communities such as foreigners or migrant workers as the most prevalent. For instance, one interviewee said, "*I think the biggest threat in India is that 80% of all the misinfo is targeted towards one particular community, that's Muslims.*" Interviewees also spoke about content that uses tactics to paint the targeted group as immoral people or to dehumanize them:

  > "*Often it'd be a claim about some horrible thing that the other group would have done—some false claim. It can often be something about how they harm women, how they harm children...accusations of rape—things that are really meant to stoke really strong emotions, and that we see pretty often.*" (P15)

  This kind of misinformation can exacerbate existing tensions, play into long-running narratives, and spill over into violence—in short, add to the other aspects already discussed. However, because the targeted group is already marginalized, the importance of addressing this particular kind of misinformation can be heightened for fact-checkers.

- **Misinformation targeted to susceptible groups.** On the flip side of misinformation that targets specific groups is when a group is targeted to be the recipients of misinformation. When these groups are targeted – a concern for most of our interviewees – the goal may be to trap them in financial scams, sell them specific products, or otherwise manipulate them into some action desired by the misinformation poster. Targeted groups for these purposes include people who may be more susceptible have low media literacy or a lack of resources or motivation in seeking credible information. Thus, members of these groups will not investigate the source or even expect a good source. Other times, susceptible groups are those who have greater fear of the unknown due to being more sheltered or who are generally vulnerable members of society. These were deemed by one interviewee to encompass:
  "*Unemployed people, elder[ly] people, more generally people who live in fear of their security. So mostly people who live outside of inner cities, people living in the suburbs, in the countryside...*" (P21)
  [1mm]
- **Misinformation aligning with existing cultural or political biases.** Finally, some interviewees talked about the believability of certain claims when it conforms with existing biases. For instance, one fact-checker felt that on the topic of politics, biases about one's opposing side may make many claims seem plausible despite not having evidence:
  "*Sometimes it's utterly believable such as when this happens a lot. The opposition party leader in India, Rahul Gandhi, is portrayed as having said a lot of things that he would not have said. Especially if I'm not positively inclined towards Rahul Gandhi, I will believe pretty much anything that I see so it's highly believable.*" Misinformation posters and spreaders, some interviewees speculated, also may have some self-interest driving their actions. People might fabricate misinformation to support a cause, politicians might amplify misinformation for their own political gains, and individuals might even spread misinformation supporting a certain belief as a means to gain community recognition.

- **National differences in preferred media formats.**We also noticed a difference in how fact-checkers from different countries talked about what characteristics help cause a piece of content to go viral. According to a U.S. fact-checker (P11), the kinds of media formats that are most likely to go viral are short text and images. This is because they are convenient to read and can easily catch people's attention.
  However, some of the fact-checkers from India offered a different opinion. From their point of view, videos are usually more influential and provocative than images or text, and therefore, the most effective media format to spread misinformation. One of the Indian fact-checkers shared his thoughts on why video is such a popular media format in India:
  "*India has a lot of people, and Indians have a lot of time, and they need to keep themselves occupied. Many of them also are unemployed right now. The economy isn't doing too well, and a great way to keep yourself occupied is consuming any and all content you can find on the Internet.*" (P4)
  As we can see, fact-checkers are sensitive to issues that are grounded in the communities and language that misinformation operates, thus highlighting how local factors seem to contribute to the spreading of rumors in different ways. As a result, a piece of misinformation designed to go viral in the U.S. may not perform well at all in a different country such as India.

3.2.5 Summary. In sum, multiple considerations inform the approach that fact-checkers take when it comes to prioritizing which claims to fact-check, much of it involving some consideration around harm. Fact-checkers try to assess possible kinds of harmful impacts of misinformation, alongside the different characteristics of the content itself that seem to increase urgency, like emotional language. Fact-checkers also take into account the larger contexts that surround misinformation: not just within the body of internet narratives or current events, but also how specific misinformation can affect local communities and countries.

3.3 RQ2: How do fact-checkers decide what to fact-check and what tools could improve their processes of prioritization?

Having understood what aspects of misinformation create a sense of urgency or importance for fact-checkers, we now inquire more about how they in fact decide what pieces of information to fact-check, and whether tools could improve their decision-making processes.

3.3.1 Fact-checking has limited capacity. We asked our interviewees whether they were able to fact-check and publish everything they thought was important, and the answer was overwhelmingly 'no.' Fact-checking organizations have limited resources and capacity, particularly if they are a small team. Some claims and conspiracy theories are inherently not fact-checkable and are avoided as a time sink. Chasing down evidence or getting a comment from a primary source can be a time-consuming task and in some cases futile:

> "*We, as a Ukrainian small organization, are not as strong an organization to ask people from [a large international corporation], 'Please give us a comment.' I wanted to, I texted them, but all this process of big companies, you text, ask their press office to give a reply, but no one gives you a reply.*" (P2)

Fact-checkers face competition from other fact-checkers in some countries, as well as time pressure to fact-check quickly in order to have the most impact. Delivering fact-checks to the right audience at the right time is as important and even as challenging as fact-checking itself:

> "*Within last week, we felt really pressured to fact-check stuff very quickly because the protests were moving at a very fast pace. If we published stuff related to last week this week, it wouldn't be timely anymore.*" (P5)

In light of limited resources, all interviewees acknowledged the necessity of making decisions on prioritizing certain claims to fact-check given competing pressures and current events, even if it meant disregarding other instances of noticed misinformation.

3.3.2 Multiple considerations go into decision-making regarding prioritization. According to our participants, there were three factors that most agree are important when deciding which piece of claims to fact-check first. The first, which is the main focus of this work, relate to aspects of urgency and importance of the claim, which interviewees described as related to harm, virality, and impact. However, interviewees also mentioned two other factors that were also important to their decision-making: limited resource allocation, as experienced around the scope of the claim, and strategic considerations due to the interests of other stakeholders.

- **Urgency of the Claim**: As revealed in subsection 3.2, interviewees validated our hunch that fact-checkers do prioritize what is urgent to fact-check based on a harms analysis, which typically incorporates the virality or reach of the claim, the possibility for negative consequences (particularly physical harm), and the potential impact the misinformation will have on society:

> "*[We triage misinformation] by how widespread the claim is. So if lots of people are going to act on it, even a small harm can get amplified; and also by how harmful it is for each individual person to act on it. So a claim that tells someone to take an untested medical remedy might be very harmful to everyone who tries it because they take something poisonous.*" (P1)

Our interviewees also reported facing difficulties in estimating the potential harm and impact of a piece of misinformation; for example certain app designs, such as E2E messaging apps, create barriers for estimating virality.

- **Resource Allocation and Claim Scope**: Interviewees also described taking into account their own limited resources and capabilities, as well as the role of fact-checking organizations and their relative strengths compared to other kinds of news or information sources. In terms of scope, fact-checkers usually prefer prioritizing claims that can be clearly refuted in a timely manner with real evidence. As a result, claims that are too hard to validate such as rumors, outlandish conspiracy theories, opinions, predictions about the future, or vague statements are all considered out of scope:

  > "*...If someone says, 'I think that Biden is a better president than Trump', we can't fact-check that...If someone says, 'By 2050 the ice capsule will melt', we can't check that. And also we can't really check very vague things, so if someone just says like, 'A lot of people think, I don't know, that guns should be banned', you can't fact-check that because what is a lot of people? You need to acknowledge your own limitations, so what we really look for when we look for things to fact-check is something very specific...*" (P5)

  Beyond claims that are out-of-scope, if a claim takes too long to fact-check or requires too many resources, fact-checkers tend to move on to claims that are easier to fact-check and come back to the "hard" claims only after the other ones have been handled.

- **Interests of Different Stakeholders**: Finally, fact-checkers often have to weigh the (potentially competing) interests of different stakeholders before deciding on which claims to fact-check, e.g., their own organizational interests and financial incentives, needs of media platforms with whom they are collaborating, the public's interest in the topic or its newsworthiness. For instance, several interviewees mentioned partnerships their organization had with social media platforms (e.g., Facebook, WhatsApp, TikTok), where those platforms had content that they wanted to prioritize for fact-checking. Some interviewees working with certain platforms talked about a pressure to write as many fact-check pieces as possible as they were paid according to how many they complete. In terms of other financial incentives, some interviewees also talked about a pressure to fact-check items that are already being widely discussed, which would then drive more traffic to the fact-check site:

  > "*I think the editors would love it if we were constantly fact-checking things that were front page news, but it also means that there's a bit of a pressure to focus on things that are maybe already getting widely spread and widely shared and talked about.*" (P1)

3.3.3 No established systematic approach towards harm assessment or claim prioritization. Finally, we found that across our interviewees, few described any kind of established or systematic method they or their organization employed to assess the potential harm associated with a piece of misinformation and or otherwise prioritize the most harmful claims. Current practices of claim selection tend to be ad-hoc and time-consuming. It is often also a collaborative decision-making process involving the entire team of fact-checkers, editors, and sometimes media partners.

> "*And when we find something valuable, something that's spread really quickly. And that reflect a topic that is important now. We discussed it between our team members. We have different channels where we discuss the things, like in Messenger or in our group on Facebook. And then we decide, should we take it or not?*" (P16)

In some cases, interviewees described disagreements between members of the team where they would debate what to cover. In other cases, individuals got to choose what they wanted to fact-check but might get suggestions from other team members or get advice from a more senior member:

> "*So first, we have a stage, that's called the media scan, where we choose our topics and...[the] whole team can give us suggestions what to choose...Typically, [one] chooses the subject he wants to write about, but sometimes he's advised by, for example, me or some other editor regarding, for example, whether we fact-check this article...from the view of fact-checker, they would most likely ask someone more senior for their opinion, which to fact-check.*" (P17)

3.3.4 Tools evaluating potential harmful impacts of misinformation. Our participants mentioned the usefulness of evaluation tools that could help them understand the potential impact of a piece of misinformation. Interviewees hoped, for example, that automated semantic analysis or detecting keywords that are related to hate speech or provocations of violence could help them assess the degree of potential harmfulness. Some interviewees also told us that auto-extracting the topics covered in a piece of content and evaluating their relevancy to local events would significantly speed up their process to select more urgent claims to fact-check.

> "*I would like it to take care of the urgency of issues in society. For instance, I would like such a system to weigh COVID-19 misinformation and political misinformation more heavily. And that's considering my own environment. So maybe in the US, it could be something else that is more urgent.*" (P19)

In addition, a few participants pointed out that such a computational tool should not only display a score or level of harm, but also show an analysis on why it is harmful:

> "*I think it would be useful to have almost at a glance, not just an understanding of what is the most harmful misinformation, but also why... It's good to be able to look at something and see, oh, this is going to be harmful because it's widely shared, and this is going to be harmful because it has a very direct impact, that helps with the prioritization.*" (P1)

Thus, fact-checkers desired to more quickly and fully understand a range of potential harmful impacts based on differing aspects, whether direct threats of physical violence to a piece of misinformation's affinity for the local context.

3.3.5 Better tools to track and measure current and potential spread of misinformation. The easiest and most direct way to estimate a claim's virality is to measure the current spread of the claim. Most interviewees reported that they found tools that show the number of viewers, the share speed, and the reach & shares of a claim across demographics useful. And for fact-checkers in international organizations, they pointed out that there are yet no tools (though demanded) that could display a global view of misinformation spread.

Fact-checkers also pointed that detecting the variations and adaptations of a claim across different platforms or media formats could also be a strong indicator of how popular a claim is. If a software can tell that the same message has been adapted to both text and video, and it is spreading in Twitter, Facebook, and other media platforms, then the fact-checkers could be fairly confident that such a claim is truly trending at the moment:

> "*Maybe something that could link narratives, like this story looks like this on Twitter; it has also spread on Telegram and it looks like this. So to be able to connect similar stories throughout the internet and have it visualize for every piece of misinformation I should fact-check. That would really help with monitoring things that has gone viral.*" (P14)

In addition to measuring the current spread of a piece of misinformation, a robust metric around virality would also aim to predict the spread of misinformation. Detecting and flagging claims with the potential to go viral, and before the public has been been widely exposed, would greatly assist fact-checking organizations.

Predicting the potential spread of misinformation require an estimation regarding the factors that make up virality, around which fact-checkers had varying thoughts. For instance, some fact-checkers believed that tools accounting for the popularity of the sender would give them a good sense of how well the claim would spread; others thought evaluating the believability of a claim would be a good signal. The majority of our interviewees desired for computational tools that could flag a content by its sentiment level or the use of specific words. Finally, because misinformation has its own life cycle, understanding where a claim is within that cycle would help the fact-checkers better evaluate the future spread of the claim:

> "*It would have to measure not only the virality, but at which point are we in the viralization? So if it's starting to viralize, if it's already past its peak point then it's coming down, or if it was already very viral but now isn't anymore.*" (P10)

3.3.6 Tools that better address local context. Many participants expressed concerns about using automated tools for selecting and evaluating claims since most of the contemporary tools do not take local knowledge into account. Fact-checking organizations usually have local experts working on the ground to provide context to understand the misinformation spreading locally. An interviewee from France told us that a computational tool can only provide substantial value if its analysis is comparable to that of a local expert:

> "*For the claim itself, I think it really does take human with local knowledge to understand how important it is. Sometimes, like I said with our fact-checkers on the ground, they might understand the importance of something more than... Even if it's in English and I can read it and understand it, they might understand more of the context and understand why it's important than some other people in another country. I'm not sure there's a tool that could really do that better than local knowledge.*" (P15)

Participants also pointed out that it would be helpful to have tools that reveal the differences in a claim's reach and shares among different demographics, as well as any targeted groups/communities if relevant. Thus, if a rumor reached a particular vulnerable group, the tool could help them know that it should be prioritized.

> "*It would be a tool that says, for example, this category of persons would be more vulnerable to questions about migration or health or safety. If the misinformation reached the category of people, who are more vulnerable to that kind of topic, then we could prioritize it.*" (P7)

But if a false claim did not have much reach, perhaps the fact-checking group would hold off from fact-checking; fact-checkers explained that giving more visibility to the misinformation when it has low reach was counterproductive.

3.3.7 Tools for collective decision-making on prioritization. Furthermore, several fact-checkers pointed out the need for tools that facilitate transparent and democratic team decision-making. Areas that a tool could assist included broadening the possibilities for participants in decision-making in ways that permitted a discussion of reasons behind certain decisions:

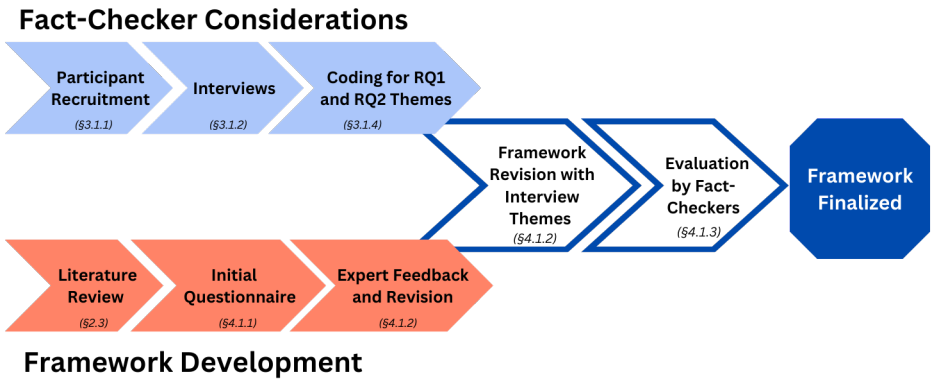

Fig. 1. Methodological Process for the development of the FABLE Framework of Misinformation Harms.

> "*I think any tool should allow us to make decisions in a more transparent way. And in that sense, if more people are involved, everyone can see what the process is, why decisions are made in a certain way, and everyone can contribute to the discussion. That would be much better than having one person centralize the decision-making.*" (P10)

In addition, one interviewee cited some concern around the potential subjectivity of decisions, especially if the power of those decisions are concentrated in one individual within a team or organization:

> "*I read about the [...] research [where someone] asked the editor how he determines which news reports go out, and which ones don't go out. And eventually it was determined that the decision-making for this editor is usually subjective, not necessarily guided by any journalistic approach. So the number one thing I would like to see in this tool is a cleansing of any form of bias that could come into fact-checking and editing. For me, as an editor. I would like to see that.*" (P19)

3.3.8 Summary. In sum, we validated that fact-checkers do make decisions regularly about prioritization. These decisions often reference signals or concepts that relate in some way to harm. Signals that fact-checkers consider include virality or spread, which affects the number of people who are exposed to the claim and may be harmed, and the possible negative consequences of believing certain claims, which can directly harm individuals. These concepts form the basis of our exploration into a harms model for misinformation. We also note that fact-checkers have additional organizational and strategic considerations not covered by our work to develop a harms model of prioritization. Finally, we find that fact-checkers do not have an established process, suggesting that tools to support structured consideration of harm could be valuable.

4 MISINFORMATION HARMS FRAMEWORK FOR PRIORITIZATION

Based on our interview results, we learn that gaining precision about the relative harmfulness of a piece of misinformation can improve the fact-checking process (and experience) of prioritization, by making it less dependent upon individual and ad hoc judgment. In addition, misinformation harm assessment can help by more clearly articulating the kinds of impact that fact-checkers seek through their work, and provide opportunities through which their work may be strategically made easier or evaluated through supporting tools.

	Fragmentation	Actionability	Believability	Likelihood of Spread	Exploitativeness
Agrafiotis et al. [5, 6]	Societal harm; Indirect harm; Long-term harm; Community- and Nation-level impacts	Physical harm; Direct harm; Short-term harm	Community-level impacts		Psychological and Emotional harm (fear); Group identity
FullFact [1]	Polarization	Gravity (or severity of harm)		Scale	Demographics; Novelty
Scheuerman et al. [41]	Coordinating Scams and Political Attacks	Coordinated Attacks (x2); Targeted Attacks; Hate speech			
Wardle & Derakhshan [51–53]			Imposter content; Fabricated content		

Table 2. Aspects of existing frameworks and taxonomies that map to our proposed framework of five dimensions of misinformation harms that convey greater urgency.

A definition and evaluation framework expressed as a urgency checklist or questionnaire around misinformation as a harm is an example of a mechanism that both helps fact-checkers prioritize their efforts while also laying the foundation upon which impact could be assessed. This work can also serve as annotation, training, or evaluation guidelines for future work in automated harm assessment support. We present the results of our effort to develop a more structured and systematic misinformation harms framework below.

4.1 Method

To lay the groundwork for this framework, we looked first to researchers who in recent years have developed taxonomies and definitions regarding harm and misinformation; these examples supported our effort to identify harmful information and misinformation needed to establish prioritization. Then, 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. Our final steps included incorporating information from our interviewees into the questionnaire, and then also gaining feedback from them and other fact-checkers(Figure 1).

4.1.1 Building from existing taxonomies. Several types of work we referenced include general harm-oriented taxonomies, misinformation-oriented taxonomies, and urgency-related taxonomies. The fuller literature review can be found in Section 2. A few of these works proved instrumental to our thinking as we clarified our own dimensions and are listed in Table 2.

For example, from harm-oriented literature, work from Agrafiotis et al. 2016 and 2018 [5, 6] defined dimensions of harm which included types such as *physical*, *economic*, and *psychological* harms, and about the broadness of impact from *individual* to *national* to *social* levels. The works

#	Gender	Country	Role	Format
R1	Male	Italy	Editor	Interview
R2	Female	France	Editor/Journalist	Survey
R3	Male	France	Fact-checker	Survey
R4	Male	South Africa	Fact-checker	Survey

Table 3. Demographic details of reviewers providing questionnaire feedback.

also considered the dimension of time, or *short-* versus *long-term* harms. These categories of harm informed four out of five of our dimensions and helped us confirm that all these dimensions of harm are represented in some part of our framework.

In another example, work led by the non-governmental organization FullFact sought to describe indicators that could help groups determine whether or not urgent coordinated action might be warranted. One of their papers described how aspects of *Scale*, *Demographics*, *Novelty*, *Polarization*, *Gravity* (or severity of harm) might help indicate the urgency around the need for action or support [1]. Similar to the Agrafiotis framework, we used this framework to confirm that a different four of five documented aspects of urgency have representation in our framework.

Scheuerman et al. brought together severity-related concerns with a harms-oriented approach to re-classify a number of harm categories provided by Facebook [41]. However, few of the categories intersected directly with the topic of misinformation; the ones that were most related are listed in the table. Overall, the paper also provided helpful insights with regards to the effort around severity-related classification itself.

Last, for a misinformation-oriented approach to taxonomization, Wardle and Derakshan's categories of *Imposter content* and *Fabricated content* [51–53] were especially noteworthy to us, as they demonstrating clear cases when people's belief and trust are potentially being harmed.

When reviewing this literature, we found that efforts sought to distinguish or characterize misinformation and harm along two different lines in ways that do not overlap. First, there is a focus upon defining different general types or *categories* of harm, e.g., physical harm, individual harm. At the same time, other efforts capture *variable magnitudes* of harm whose value may vary according to context. For example, the reach or 'virality' of a piece of misinformation can have a value that is higher or lower depending on factors such as the popularity of the poster. When characterizing misinformation as a harm in order to assess urgency, clarifying the difference between these *categorical and variable* characteristics can be useful. It is clear among variable characteristics when a case is more urgent, for example, when something has more virality as opposed to less. However, there may always be disagreements and considerations among which categories of harm matter more: are issues with short-term impacts always more important to address compared to those with long-term ones? Thus, it is easier to define degrees of urgency within a category of harm as compared to across categories.

4.1.2 Taxonomy and questionnaire development. In addition to reviewing and incorporating thematic concepts for evaluating harm that we discovered in our review of existing taxonomies, we incorporated insights from our own past engagements with fact-checkers. Through informal workshops and conversations with other experts, we also received feedback as we iterated on our taxonomy and set of questions as a team over the course of about a year.

As we conducted and analyzed the interviews we conducted with fact-checkers, we began to incorporating the themes into the taxonomy. Their considerations regarding the harmfulness of misinformation not only validated our approach, but suggested additional areas for consideration.

Fig. 2. Dimensions of the FABLE Framework of Misinformation Harms.

For instance, the creation of a dimension that we call "social fragmentation," which addresses longer-term community or societal level harms, resulted specifically from their observations addressed in Section 3.3.1 and 3.3.4.

4.1.3 Evaluation. As the framework became more finalized, we turned towards a more formal feedback process. In our final step, we invited four experts in misinformation and fact-checking to work through our questionnaire and share their feedback (Table 3). With each expert, we either conducted a 30-minute interview or asked them to complete a survey which guides them through the questionnaire and prompts them to provide feedback on each of the five dimensions.

All reviewers agreed that our framework matches with their own understanding of misinformation harm and could prove useful in multiple scenarios. R4 mentioned that "*A framework like this would be very useful, and my organization already uses an informal, unwritten version of this kind of framework.*" Similarly, while R3 pointed out that he himself wouldn't need to use the framework as he already had abundant experience in defining harm, he thinks the framework "*could definitely be useful during teaching sessions or for younger journalists.*"

At the same time, reviewers offered insightful suggestions on the questions' coverage, readability, and conciseness, which we incorporated to further refine our framework. For instance, R1 requested greater clarification about narrative on a question within our "Social Fragmentation" dimension that originally stated "*Does the message fit into a larger narrative that has been existing for some time?*" In response, we modified the question. Using the work of psychologist Jerome Bruner on narrative[13], the new version reads: "*Does the message fit into a larger story or argument, for example about how the world works or how people think?*" with a note explaining its significance: "*A larger narrative may include stories about communities, race, poltitical parties. A larger narrative crosses platforms and has existed for some time.*"

4.2 FABLE: A Five-Dimension Urgency Framework for Assessing Misinformation Harms

We present a framework of five variable "dimensions of urgency," along with a questionnaire, that can support fact checkers in their efforts to discuss and prioritize in a more strategic way (Figure 2). These five dimensions can help to clarify urgency to address a piece of misinformation, thereby helping response teams have a clearer understanding of what impacts they may want to achieve. It may be that certain organizations prioritize certain classes of fact-checking content. But by providing multiple dimensions, fact-checkers have a view to a holistic approach that reveals what issues might be missed if organizations are always only focused on one or a few dimensions.

Our "FABLE Framework" is a model includes five dimensions of urgency when it comes to assessing and prioritizing misinformation as a harm: **(social) Fragmentation**, **Actionability**, **Believability**, **Likelihood of spread**, and **Exploitativeness**. The five dimensions of urgency are each defined through a set of questions, which are incorporated into a single multi-part

questionnaire (see Appendix A and online resource[2]); the questions that make up the questionnaire reflect factors that are currently understood to have an impact upon misinformation's magnitude or potential harm.

In the following, we define each dimension according to the relationship between misinformation and potential harm. We provide key questions associated with each dimension as well. Appendix A has the full list of questions associated with each dimension along with tips for and examples of what to look for.

4.2.1 *Dimension: (Social) Fragmentation.*

Definition: 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.

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 a long-term consequence of misinformation is reduced trust in existing institutions and social groups.

This category emerged out of our exchanges with fact-checkers and work on a related project. Many of these questions about trust in themselves may be important for functioning societies and democracies to work. Indeed, scholars have argued that a certain amount of distrust may be necessary, for example [23]. Hard questions about the appropriate and just functioning of public institutions, the scientific community, or the media are appropriate to ask. However, our framework is focused on the effect of repeated questions such as these when combined with incorrect or inaccurate information.

Questions include:

- Does the message fit into a larger story or argument, for example about how the world works or how people think?
- 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?

4.2.2 *Dimension: Actionability.*

Definition: 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.

Questions tied to this dimension are intended to ascertain whether characteristics or factors related to the message(s) 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 this call can be obscured. We focus on key questions for this dimension on more overt signals regarding direct action

[2]https://www.artt.cs.washington.edu/analysis-framework-online-misinformation-harm

or coordination; additional questions listed in Appendix A attempt to capture subtler considerations that may heighten actionability. These include messages that cast aspersions on particular groups, make use of injustice or moral outrage, or are tailored or addressed to communities with a history of violence.

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

4.2.3 *Dimension: Believability.*

Definition: 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.

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.

Key questions focused on either the inability of readers to easily verify information, whether strong communities or audiences already exist around for certain topics, or whether publishers have unclear editorial practices. Other questions take note of the absence of corroborating evidence around certain issues as well as the familiar tone of the messages. The issue of imposter content or accounts are important to this category, as belief is achieved by taking advantage of a community's good will.

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)?

4.2.4 *Dimension: Likelihood of Spread.*

Definition: 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.

These questions try to ascertain whether characteristics or factors related to the message(s) make the content likely to spread or be discovered. 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.

Our key questions focus on the accounts or persons with histories with large reach or repeated instances of advancing rumors. Other questions go beyond to look at aspects of information context, such as platform design, and characteristics of the content itself, such as direct appeals to the audience or its format (e.g. text versus audio or video). We also included considerations of current events and novel trends, as well as the tone of the message.

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?

4.2.5 *Dimension: Exploitativeness.*

Definition: 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.

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 and the elderly are related but focus on characterizing the group, whereas this dimension strives to examine when aspects of the message itself directly engage in exploitation.

Our key questions highlight common strategies for exploiting particular weaknesses or vulnerable groups. The remaining questions in Appendix A expand to ask about additional vulnerable groups, such as veterans or conspiracy theorists, or other kinds of weaknesses, such as feelings of isolation, powerlessness, or disenfranchisement. We also consider whether the content is in a less popular language, where there are fewer protections and resources for fact-checking, or is spreading in a region where local audiences may be more vulnerable.

Key questions for this dimension 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 topic or explanation complicated?

4.3 Case Study

Our five-dimension FABLE framework attempts to improve the ability for fact-checkers to evaluate the potential harmfulness of an inaccurate or misleading claim and prioritize it as deserving of attention from their organizations. This, however, does not result in a single, quantifiable ranking of urgency. Rather, the framework takes into consideration the many characteristics of misinformation that may lend themselves to harmful impact outlined under RQ1. It then clarifies the

potential magnitude of harm based on *variable* dimensions, while leaving open the possibility that, for strategic purposes, a fact-checker or organization may choose to focus on different categories of harm.

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.

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, excepting imminent physical danger, 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 it may be possible to clarify which cases might be more urgent among similar cases, as well as to help distinguish where more strategic decision making needs take place. This clarification can help both prioritization and impact assessment.

Answering the questions connected to the framework, a fact-checker may realize that neither scenario seems likely to indicate *actionability* and both rank highly in the *likelihood of spread*. However, the former scenario may tally higher in the area of *believability*, while the latter results in some concerns about *social fragmentation*. Depending on how the fact-checker and their organization decide to weigh these dimensions, as well as how they might approach more categorical considerations such as the potential for broad physical impact (e.g. on female bodies) versus individual reputational and social harms, the decision for ultimate prioritization between these two claims may be different.

5 DISCUSSION

We discuss the implications of our results for practitioners and for researchers. From the interviews, we learned that fact-checkers are clearly concerned about harm, and that a focus on misinformation impact can be a productive avenue for investigation. Whether for practice or for research, defining the difference between categorical and variable dimensions better clarifies key decisions related to policy and organizational strategy. In other words, the prioritization of categorical harms, such as physical versus psychological damage, is one that has to be made by humans, and cannot be determined by automated solutions. This recognition, as expressed in the framework itself, leads to a number of implications for both practice and research.

Implications for practitioners. For practitioners in fact-checking organizations, the framework is one solution to the expressed needs by our interviewees for more systematic approaches to prioritization. Going through the questionnaires with several examples of claims can help organizations decide which kinds of categorical and variable dimensions they want to prioritize, thereby allowing them to customize the framework to their own needs.

Our expert reviewers saw the benefits of the more nuanced structured assessment of our framework. In fact, R2 mentioned that different dimensions could be weighted differently, and cautioned us against adding the scores of each dimension together to form a conclusion or make comparisons. There may be value in using the framework in a periodic way for organizations seeking to structure their decision-making and increase their internal transparency, an expressed desire mentioned under RQ2. Moreover, R1 emphasized that since the framework provides a comprehensive breakdown

of the various aspects of harm, it could be an invaluable tool for training new fact-checkers and editors.

However, the initial application of the framework is not quick. Though answering these questions becomes faster with practice, as discussed under RQ2, fact-checking organizations are limited in their time, capacity, and resources. Finding efficient ways to apply this framework to daily prioritization remains an area to be further studied.

Implications for researchers. The need for efficiency creates opportunities for researchers. As in past research and this paper, the complexities of harm and misinformation are high, which points to an area where HCI solutions can support structured, thoughtful, and faster approaches by fact checkers.

The different dimensions of the framework are conceptual tools for fact-checkers to negotiate their own processes. They also create design opportunities that give more decision-making control back to the users/organizations: rather than dictating that a piece of misinformation is more urgent the more viral it is, for example, applying the taxonomy lends itself towards interesting HCI possibilities, such as providing metrics in dashboards and visualizations that qualitatively and quantitatively represent the impact of their work. Additionally, HCI solutions that seek to integrate the taxonomy within a single fact-checking organization might result in platforms where team members can negotiate their processes or discuss their biases and impact as an organization.

Additionally, a clear area of work is the potential automation of the taxonomy towards a prioritization queue or as a filtering and triage system—desires expressed by interviewees in section 3.3. Automated misinformation prioritization has been extremely challenging due to the complexities and nuances of misinformation harm. With the recent advancements in Large Language Models (LLMs), our framework offers a pathway to guide and fine-tune these models, enabling them to generate structured and explainable analyses of harm. Most of the reviewers (R1, R2, and R4) pointed out that the framework could be immensely beneficial if developed into a future automated tool for the preliminary screening, filtering, and prioritization of harmful content. For example, R4 responded "*This would absolutely be useful. A major issue (if not THE major issue) with existing misinformation-detection tools is that they perform extremely poorly at prioritizing false claims.*" And R3 asserted that "*having a tool to detect harmful content and develop filters could be a way to prevent from post-traumatic stress disorder for example.*" In addition, a new research agenda for what indices might be automated, such as actionability or likelihood of spread, are clear opportunities for saving time and effort.

6 LIMITATIONS

While our effort aligns with the importance that professional fact-checkers place upon the potential harms of misinformation, we learned through our interviews that this was not their only concern. They also negotiate the needs of competing stakeholders, as well as their deliverables for tech platforms. Our effort however only focuses on harm. How these other priorities might be taken into account is an area for future research. Additionally, as mentioned earlier, we chose to focus upon professionalized fact-checking processes. However, developments in collaborative, crowdsourced contexts will offer additional ways to consider the problem of prioritization which may be beneficial for professional contexts.

Related to the specifics of the framework, we note that the validation by our reviewers is not robust. Rather, this evaluation served as an initial confirmation of our findings; more investigation of the framework, and its possible transformation, is work that is yet needed. Also, 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.

7 FUTURE WORK

In addition to building upon the framework to investigate which aspects or elements could be designed and automated, and therefore more quickly assist fact-checkers, other kinds of assistance could also be explored based upon the comments of our interviewees.

Filtering for verifiable claims. As pointed out by our fact-checkers, not all claims are fact-checkable. Many comments or conspiracy theories with potential for harm cannot be fact-checked or debunked. Tools that could separate and identify fact-checkable claims from noise, or that could isolate what elements of a claim are able to check, would prevent fact-checkers from spending time on information that is inherently non-verifiable.

Automatic format processing and automatic clustering. Most of our interviewees also reported that watching videos was an extremely time-consuming task in their fact-checking process. They pointed out the need for video processing tools that automatically summarizes the important messages conveyed in the video into text, or group videos together by their topics or themes.

In addition, some fact-checkers told us that they maintain collections of fact-checked claims and their results. For instance, a Brazilian fact-checker we interviewed claimed that having a tool that would automatically match new claims to the existing claims in their database and send out the results would hugely improve their fact-checking efficiency.

Resources in less popular languages and contexts. Lastly, an Ukrainian fact-checker reminded us of the inequality in the available resources between countries and languages. Most NLP models are trained on popular languages such as English and Spanish, while few tools are available for people that speaks a minor language such as Ukrainian. In future development of AI or NLP, we should be reminded that these computational tools need to be language inclusive.

8 CONCLUSION

In this work, we examined how fact-checkers prioritize which claims to inspect for further investigation and publishing, and what tools may assist them in their efforts. We explored this through interviews with 23 fact-checkers around the world, clarifying what aspects of misinformation they considered to create urgency or importance. These often revolved around the potential for the claim to harm others. We also learn more about the processes behind fact-checking decisions and suggestions for tools that could help fact-checkers with them. To address the needs articulated by these fact-checkers and others, as well as a gap in the framework literature, we present the FABLE Framework of Misinformation Harms: a five-dimension questionnaire to help fact-checkers negotiate the priority of claims. This effort was further validated by additional interviews with expert fact-checkers.

9 ACKNOWLEDGMENTS

We appreciate the input of our fact-checkers and experts over the past two years, 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 at the University of Washington for their feedback. We very much thank key contributions of Skyler Hallinan and Alexandra Bornhoft to this paper as well. Some of this work was supported by a larger project funded by the National Science Foundation's Convergence Accelerator program under Award No. 49100421C0037.

REFERENCES

[1] 2020. *A framework for misinformation crises - Paper 1: identifying scope and risk.* Technical Report. FullFact. https://fullfact.org/media/uploads/draft_paper_1__identifying_scope_18_november_2020.pdf

[2] 2020. Twelve taken ill after consuming 'coronavirus shaped' datura seeds. *The Hindu* (Apr 2020). https://www.thehindu.com/news/national/andhra-pradesh/twelve-taken-ill-after-consuming-coronavirus-shaped-datura-seeds/article31282688.ece

[3] 2021. Transparency Report 2021 - Reddit. https://www.redditinc.com/policies/transparency-report-2021-2/

[4] n.d.. YouTube Community Guidelines enforcement – Google Transparency Report. https://transparencyreport.google.com/youtube-policy/removals?hl=en

[5] Ioannis Agrafiotis, Maria Bada, Paul Cornish, Sadie Creese, Eva Ignatuschtschenko, Taylor Roberts, and David Upton. 2016. *Cyber Harm: Concepts, Taxonomy and Measurement.* Technical Report RP 2016-23. University of Oxford. http://www.ssrn.com/abstract=2828646

[6] Ioannis Agrafiotis, Jason R C Nurse, Michael Goldsmith, Sadie Creese, and David Upton. 2018. A taxonomy of cyber-harms: Defining the impacts of cyber-attacks and understanding how they propagate. *Journal of Cybersecurity* 4, tyy006 (Jan. 2018). https://doi.org/10.1093/cybsec/tyy006

[7] Paul M. Barrett. 2020. *Who Moderates the Social Media Giants? A Call to End Outsourcing.* https://issuu.com/nyusterncenterforbusinessandhumanri/docs/nyu_content_moderation_report_final_version/1

[8] Paul Bleakley. 2023. Panic, pizza and mainstreaming the alt-right: A social media analysis of Pizzagate and the rise of the QAnon conspiracy. *Current Sociology* 71, 3 (May 2023), 509–525. https://doi.org/10.1177/00113921211034896

[9] Sissela Bok. 2011. *Lying: Moral Choice in Public and Private Life.* Knopf Doubleday Publishing Group. Google-Books-ID: F3ySLLhv7LMC.

[10] Elena Bozzola, Giulia Spina, Rocco Russo, Mauro Bozzola, Giovanni Corsello, and Alberto Villani. 2018. Mandatory vaccinations in European countries, undocumented information, false news and the impact on vaccination uptake: the position of the Italian pediatric society. *Italian Journal of Pediatrics* 44, 1 (Jun 2018), 67. https://doi.org/10.1186/s13052-018-0504-y

[11] J Scott Brennen, Felix M Simon, Philip N Howard, and Rasmus Kleis Nielsen. 2020. *Types, sources, and claims of COVID-19 misinformation.* Ph. D. Dissertation. University of Oxford.

[12] Corey L. Brettschneider. 2012. *When the State Speaks, What Should it Say?* Princeton University Press, Princeton, NJ and Oxford, UK.

[13] Jerome Bruner. 2010. Narrative, Culture and Mind. In *Telling stories: language, narrative, and social life*, Deborah Schiffrin, Anna De Fina, and Anastasia Nylund (Eds.). Georgetown University Press, Washington, DC, 45–49.

[14] Marina Charquero-Ballester, Jessica G Walter, Ida A Nissen, and Anja Bechmann. 2021. Different types of COVID-19 misinformation have different emotional valence on Twitter. *Big Data & Society* 8, 2 (2021), 20539517211041279.

[15] Johan Farkas and Jannick Schou. 2019. *Post-Truth, Fake News and Democracy: Mapping the Politics of Falsehood.* Routledge. Google-Books-ID: fMuqDwAAQBAJ.

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

[17] Sara Fischer. 2020. Fact-checking goes mainstream in Trump era. *Axios* (Oct 2020). https://www.axios.com/2020/10/13/fact-checking-trump-media

[18] Mary Ann Fitzgerald. 1997. Misinformation on the Internet: Applying evaluation skills to online information. *Teacher Librarian* 24, 3 (1997), 9.

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

[20] FullFact. 2020. *Towards a framework for information incidents - Paper 3: Levels of Incidents.* Technical Report. FullFact. https://fullfact.org/media/uploads/framework_paper_3.pdf

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

[22] Lucas Graves, Brendan Nyhan, and Jason Reifler. 2016. Understanding Innovations in Journalistic Practice: A Field Experiment Examining Motivations for Fact-Checking. *Journal of Communication* 66, 1 (Feb 2016), 102–138. https://doi.org/10.1111/jcom.12198

[23] Russell Hardin. 2006. *Trust.* Polity Press.

[24] Lu He and Changyang He. 2022. Help Me #DebunkThis: Unpacking Individual and Community's Collaborative Work in Information Credibility Assessment. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–31.

[25] Shanto Iyengar and Douglas S. Massey. 2019. Scientific communication in a post-truth society. *Proceedings of the National Academy of Sciences* 116, 16 (Apr 2019), 7656–7661. https://doi.org/10.1073/pnas.1805868115

[26] Shan Jiang and Christo Wilson. 2021. Structurizing misinformation stories via rationalizing fact-checks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 617–631.

[27] Prerna Juneja and Tanushree Mitra. 2022. Human and technological infrastructures of fact-checking. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–36.

[28] Ralph Keyes. 2004. *The post-truth era: dishonesty and deception in contemporary life*. St. Martin's Press, New York.

[29] Stephan Lewandowsky, Ullrich K. H. Ecker, and John Cook. 2017. Beyond Misinformation: Understanding and Coping with the "Post-Truth" Era. *Journal of Applied Research in Memory and Cognition* 6, 4 (Dec 2017), 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008

[30] Sahil Loomba, Alexandre de Figueiredo, Simon J. Piatek, Kristen de Graaf, and Heidi J. Larson. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour* 5, 33 (Mar 2021), 337–348. https://doi.org/10.1038/s41562-021-01056-1

[31] Aaron M McCright and Riley E Dunlap. 2017. Combatting misinformation requires recognizing its types and the factors that facilitate its spread and resonance. (2017).

[32] Nicholas Micallef, Vivienne Armacost, Nasir Memon, and Sameer Patil. 2022. True or False: Studying the Work Practices of Professional Fact-Checkers. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (2022), 1–44.

[33] Shujaat Mirza, Labeeba Begum, Liang Niu, Sarah Pardo, Azza Abouzied, Paolo Papotti, and Christina Popper. 2023. Tactics, threats and targets: Modeling disinformation and its mitigation. In *NDSS 2023, Network and Distributed System Security Symposium, 27 February-3 March 2023, San Diego, California, USA*, Usenix (Ed.). San Diego. Copyright Usenix. Personal use of this material is permitted. The definitive version of this paper was published in NDSS 2023, Network and Distributed System Security Symposium, 27 February-3 March 2023, San Diego, California, USA and is available at : http://dx.doi.org/10.14722/ndss.2023.23657.

[34] Kara N Moore and James Michael Lampinen. 2016. The use of recollection rejection in the misinformation paradigm. *Applied Cognitive Psychology* 30, 6 (2016), 992–1004.

[35] Charles A Morgan III, Steven Southwick, George Steffian, Gary A Hazlett, and Elizabeth F Loftus. 2013. Misinformation can influence memory for recently experienced, highly stressful events. *International journal of law and Psychiatry* 36, 1 (2013), 11–17.

[36] Kai Nakamura, Sharon Levy, and William Yang Wang. 2019. r/Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. *CoRR* abs/1911.03854 (2019). arXiv:1911.03854 http://arxiv.org/abs/1911.03854

[37] Erik C. Nisbet, Chloe Mortenson, and Qin Li. 2021. The presumed influence of election misinformation on others reduces our own satisfaction with democracy. *Harvard Kennedy School Misinformation Review* (Mar 2021). https://doi.org/10.37016/mr-2020-59

[38] Abhishek Pandey and Alison P. Galvani. 2023. Exacerbation of measles mortality by vaccine hesitancy worldwide. *The Lancet Global Health* 11, 4 (Apr 2023), e478–e479. https://doi.org/10.1016/S2214-109X(23)00063-3

[39] Jeffrey K. Riley. 2022. Angry Enough to Riot: An Analysis of In-Group Membership, Misinformation, and Violent Rhetoric on TheDonald.win Between Election Day and Inauguration. *Social Media + Society* 8, 2 (Apr 2022), 20563051221109188. https://doi.org/10.1177/20563051221109189

[40] Joni Salminen, Hind Almerekhi, Milica Milenković, Soon-gyo Jung, Jisun An, Haewoon Kwak, and Bernard J Jansen. 2018. Anatomy of online hate: developing a taxonomy and machine learning models for identifying and classifying hate in online news media. In *Twelfth International AAAI Conference on Web and Social Media*.

[41] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R Brubaker. 2021. A framework of severity for harmful content online. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–33.

[42] Melanie Schuster, Juhani Eskola, and Philippe Duclos. 2015. Review of vaccine hesitancy: Rationale, remit and methods. *Vaccine* 33, 34 (Aug 2015), 4157–4160. https://doi.org/10.1016/j.vaccine.2015.04.035

[43] Connie Moon Sehat and Farah Lalani. 2021. *Advancing Digital Safety: A Framework to Align Global Action*. World Economic Forum. https://www3.weforum.org/docs/WEF_Advancing_Digital_Safety_A_Framework_to_Align_Global_Action_2021.pdf

[44] Sam Shead. 2020. TikTok is luring Facebook moderators to fill new trust and safety hubs. https://www.cnbc.com/2020/11/12/tiktok-luring-facebook-content-moderators.html Section: Technology.

[45] Samikshya Siwakoti, Kamya Yadav, Nicola Bariletto, Luca Zanotti, Ulas Erdogdu, and Jacob N. Shapiro. 2021. How COVID drove the evolution of fact-checking. *Harvard Kennedy School Misinformation Review* (May 2021). https://doi.org/10.37016/mr-2020-69

[46] C. R. Snyder. 2002. Hope Theory: Rainbows in the Mind. *Psychological Inquiry* 13, 4 (Oct 2002), 249–275. https://doi.org/10.1207/S15327965PLI1304_01

[47] Mark Stencel, Eric Ryan, and Joel Luther. 2022. Fact-checkers extend their global reach with 391 outlets, but growth has slowed. https://reporterslab.org/fact-checkers-extend-their-global-reach-with-391-outlets-but-growth-has-slowed/

[48] Anselm Strauss and Juliet M Corbin. 1997. *Grounded theory in practice*. Sage.
[49] Kurt Thomas, Devdatta Akhawe, Michael Bailey, Dan Boneh, Elie Bursztein, Sunny Consolvo, Nicola Dell, Zakir Durumeric, Patrick Gage Kelley, Deepak Kumar, Damon McCoy, Sarah Meiklejohn, Thomas Ristenpart, and Gianluca Stringhini. 2021. SoK: Hate, Harassment, and the Changing Landscape of Online Abuse. In *2021 IEEE Symposium on Security and Privacy (SP)*. 247–267. https://doi.org/10.1109/SP40001.2021.00028
[50] Thi Tran, Rohit Valecha, Paul Rad, and H. Raghav Rao. 2020. An Investigation of Misinformation Harms Related to Social Media during Two Humanitarian Crises. *Information Systems Frontiers* (Nov. 2020). https://doi.org/10.1007/s10796-020-10088-3
[51] Claire Wardle. 2016. https://www.cjr.org/tow_center/6_types_election_fake_news.php
[52] Claire Wardle. 2017. Fake news. It's complicated. https://firstdraftnews.org/articles/fake-news-complicated/
[53] Claire Wardle and Hossein Derakhshan. 2017. Information Disorder: Toward an interdisciplinary framework for research and policy making. https://rm.coe.int/information-disorder-report-november-2017/1680764666
[54] Taha Yasseri and Filippo Menczer. 2023. Can crowdsourcing rescue the social marketplace of ideas? *Commun. ACM* 66, 9 (2023), 42–45.
[55] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. *arXiv preprint arXiv:1902.09666* (2019).
[56] Han Zheng and Rich Ling. 2021. Drivers of social media fatigue: A systematic review. *Telematics and Informatics* 64 (Nov 2021), 101696. https://doi.org/10.1016/j.tele.2021.101696
[57] Ethan Zuckerman. 2019. QAnon and the Emergence of the Unreal. *Journal of Design and Science* 6 (Jul 2019). https://doi.org/10.21428/7808da6b.6b8a82b9

A APPENDIX A: MISINFORMATION HARMS QUESTIONNAIRE

The FABLE Framework of Misinformation Harms aims to support fact checkers in their efforts to discuss and prioritize in a more strategic way.

Its five dimensions can help to clarify urgency within a category of harmful misinformation, thereby helping response teams. These multiple dimensions 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.

In this framework, the degree of harm is positively correlated with degree of urgency: more "Yes" answers suggest more urgency or potential harm. However, this assumption may not always hold. We also recognize that organizations have considerations such as strategic areas of focus or internal resources. Therefore, organizations should use adjust the framework according to their needs. This is why there is no exact determination of how many "Yes" answers are required before a situation is considered urgent; it is up to each organization to determine. Or, while this framework aims to be holistic, an organization's use of it may also focus on a specific subset of dimensions or questions.

FABLE Framework of Misinformation Harms

A Structured Response to Misinformation as Harm

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—**social Fragmentation**, **Actionability**, **Believability**, **Likelihood of spread**, and **Exploitativeness.** A longer version of the questionnaire with helpful hints for answering can be found at: [redacted]

FABLE Framework

Dimension 1: Social Fragmentation

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

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

Suggestions for hints on how to answer these questions can be found at [redacted]

There are no short cut questions for this category.

Social Fragmentation Questions	Y	N	?
Does the message fit into a larger story or argument, for example about how the world works or how people think? *A larger narrative means that there is story touching the information that crosses platforms and time. This may include stories about communities, race, political parties.*			
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*			

Fragmentation *Dimension 1 Subtotal*	Actionability *Dimension 2 Subtotal*	Believability *Dimension 3 Subtotal*	Likelihood of Spread *Dimension 4 Subtotal*	Exploitativeness *Dimension 5 Subtotal*	TOTAL

**"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.*

FABLE Framework

Dimension 2: 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 [redacted]

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

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*			

**"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.*

FABLE Framework

Dimension 3: 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 [redacted].

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?			
Is there a lack of consensus on the part of experts regarding the claim?			
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?			
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 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*			

FABLE Framework

Dimension 4a: Likelihood of Spread

A piece of harmful content is more harmful the more places it appears on 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 [redacted].

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

FABLE Framework

Dimension 4b: Likelihood of Spread

A piece of harmful content is more harmful the more places it appears on 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 [redacted]

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*			

Don't forget the questions from page 4

FABLE Framework Dimension 5: Exploitativeness

A piece of misinformation is more harmful the more the message seeks to exploit 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 [redacted].

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?			
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 being shared by within an online group, community, or thread that has a recent history of discussing conspiracy theories or viral misinformation?			
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*			

B APPENDIX B: INTERVIEW CODEBOOK

This section contains the interview codes related to how our interview participants determine urgency and importance and the processes they take for deciding what to fact-check. When it was necessary for clarifying among our coders, additional notes are included for more clearly defining the code.

Parent Code	Child Code	Notes on the Definition (Optional)
Fact-checking has crucial real-world consequences and contains high stakes.	fact-checking has high-stakes	fact-checking has significant real-world consequences
	sensitive nature of political misinfo	conflict of interest between stakeholders; government involvement
	fact-checking to protect minority groups	
	fact-checking to improve public discourse	"fact-checking to change political discourse"
	fact-checking to change people's mind	persuade people without strong opinions help people make informed decisions. "fact-checking targets people in the middle"
	fact-checking to promote good information	"fact-checking for factual verification"
	media literacy as the goal	impossible to fact-check everything so improving the public's media literacy is the goal, but it could be a challenge
	publishing fact-checks to stop the spread of misinfo	publishing critical fact-checks is time-pressured. fact-checking to prevent harmful effects.
	collaboration with media partners	alerting media partners about circulating misinfo; media partners may influence prioritization
	fact-checking benefits journalism	
Fact-checking has limited capacity yet facing many challenges.	harm has been done before fact-checking	

Continued on next page

(Continued)

	people sharing misinfo are affected by misinfo	
	psychological harm is hard to observe	
	limited signals indicating virality	on certain social media platforms such as WhatsApp
	virality easier to measure	than impact or harm
	E2EE messaging apps create barriers in estimating virality	
	attacks on fact-checkers	personal safety issue, hate speech, common in some countries
	emotional distress in fact-checkers	
	misinfo harm/urgency dependent on local context	both direct context: things happened immediately before the incident, and indirect context: things going on in the long-term (relating to social and cultural norms). "religious/ethnic lines induce social conflict"
	reputational psychological harm are hard to undo	
	impact of fact-checking is hard to measure	
	extreme views are shared by the most active users online	
Fact-checkers weigh multiple factors when triaging misinfo. The most important ones are virality, imminence of real-world harm, and fact-checkability.	triaging misinfo based on toxicity	toxicity of the language
	triaging misinfo based on virality	how viral it is right now and how viral it potentially might be
	triaging misinfo based on fact-checkability	knowingly false but not debunkable/falsifiable; triaging misinfo based on easiness

Continued on next page

(Continued)

	triaging misinfo based on urgency	imminence of harm, direct harm, actionability (one way or both ways)
	triaging misinfo not based on audience	due to difficulty in measuring audience susceptibility
	triaging misinfo based on impact of fact-checking	would the fact-checking receive public attention and promote changes
	triaging misinfo based on impact	political impact, societal impact, harmful impact/harmfulness, scope of impact
	triaging misinfo based on competing interests	the fact-check's relevance to fact-checker's safety, public interest, political interest, and media partners' interest; may require trade-offs
	triaging misinfo based on information gap	audience is missing the information they need to make right decisions
	evaluating harm/urgency based on misinfo type	
Virality has many factors and depends on context, but the primary contributors are emotion/sensation, relevancy to current events, and time sensitivity.	convenience to view correlates to virality-harm	media format (images, text, video), design friction (blurring or attaching a warning to content), etc.; easiness to understand. "images being the most viral format"
	popularity of the platform contributes to virality	user base of the platform
	urgency contributes to virality	time sensitivity of the message
	difficulty in validation indirectly contributes to virality	the more difficult it is to validate a claim, the more time it takes, hence the more viral a claim goes. Language barrier makes validation more difficult.

Continued on next page

(Continued)

	sender of the claim contributes to virality	credibility/popularity of the sender; relationship between the sender and audience (audience's trust in the sender)
	media format not a crucial factor to virality	
	media format being a major factor of virality	
	easiness to remember contributes to virality	
	believability contributes to virality	there is a rational part and an emotional part of believability in terms of in-group and out-group believability
	threshold of virality varies by country	the more social media users, the more widespread something has to be in order to be considered viral
	easiness to understand contributes to virality	
	emotion/sensation as the (primary) factor to virality	provocative misinfo (e.g., ones targeting communities); appealing to empathy/sense of injustice, or negative emotions (fear, hate, sadness); tone of the claim matters. "emotion influences believability" "attention grabbing contributes to virality" "attractiveness contributes to virality" "relatability contributes to believability" "negative tones tend to be more viral"
	information gap contributes to virality	especially in health and science
	relevancy / current events contribute to misinfo virality	
Misinfo believability has both rational and irrational parts	believability contributes to virality	there is a rational part and an emotional part of believability in terms of in-group and out-group believability

Continued on next page

(Continued)

	believability influences people without strong opinions	
	trust of the sender influences believability	
	plausibility/ credibility influences people without strong opinions	hence its influence on virality depends on level of polarization, existing beliefs, etc. The more polarized a society is or the stronger existing beliefs are, the less influence believability is.
	believability differs among groups	in terms of political party, age groups, etc.
	believability can be hard to measure	
	confirmation bias influences believability	
	uncertainty contributes to believability	
People promote misinfo for personal interests	spreading misinfo for community recognition	for personal gains
	misinfo wielded for political gain	politicians weaponize misinfo to induce communal hate for political gain
	fabricating misinfo to support a cause	
	harm amplified by politicians	
Emotional value makes misinfo dangerous/urgent	emotion/sensation as the (primary) factor to virality	provocative misinfo (e.g., ones targeting communities); appealing to empathy/sense of injustice, or negative emotions (fear, hate, sadness); tone of the claim matters. "emotion influences believability" "attention grabbing contributes to virality" "attractiveness contributes to virality" "relatability contributes to believability" "negative tones tend to be more viral"

Continued on next page

(Continued)

	all misinfo has psychological impact	"reputational psychological harm are interrelated"; "misinfo targeting communities induces psychological harm"
	urgency contributes to misinfo virality	
	misinfo induces/exacerbates hate	induce hate toward a person or group; stoke communal tensions
	appealing to emotion makes misinfo more dangerous	
	emotion/sensation/urgency/actionability leads to short-term impact	
Physical and societal harm are the two major concerns of misinfo, and they are interrelated.	societal harm and physical harm are often intertwined	
	reputational harm may lead to physical harm	
	societal harm being the major threat	
	prevalence of health-related misinfo	
	prevalence of political misinfo	
	misinfo undermines trust in science/government	which has long-term impacts
	societal harm has long-term/indirect impacts	
	actionability contributes to urgency	"direct harm being more damaging"
	threatening democracy being the biggest threat	misinfo affects policymaking and government operation, sways voters, and induces polarization
	political misinfo has long-term impact	
	physical harm being the biggest threat	potential to evoke an action leads to physical harm

Continued on next page

(Continued)

	prevalence of social and religious impact	in regions like India and African countries
	impactful misinfo has short-term long-term harm	
	short-term direct long-term indirect harm are intertwined	
	triaging misinfo based on physical harm	potential to incite violence or induce harm to oneself
Individual pieces of misinfo form large narratives accumulate in impact	misinfo spreads across platforms, languages, and cultures	misinfo narratives spread across the globe: same or similar narratives recurring in multiple countries
	identifying debunking dominate narratives	
	misinfo that fits into larger narratives has long-term impact	
	accumulative effects of misinfo	misinfo harm accumulates with time and quantity. network effect of innocuous content
	misinfo carries narratives	pattern is repeating narratives
	misinfo overwhelms our capacity to believe	with risk of being people no longer believe anything
	misinfo undermines trust in science/government	which has long-term impacts
	misinfo propagates beliefs	and it is often used to propagate beliefs
Misinfo mostly targets communities	misinfo targets famous figures	
	misinfo targets individuals that represents a community	
	misinfo seldomly targets organizations	

Continued on next page

(Continued)

	misinfo mostly targets communities	that's probably why social and religious impact is prevalent, and why misinfo induces social conflict. political misinfo tends to target communities. toward minority groups. by sowing doubts.
	misinfo seldomly targets individuals	by falsely claiming death or impersonating/fabricating misinfo for profit
	fact-checking to protect minority groups	
	misinfo targets organizations	by spreading conspiracy theory
People with low media literacy are more susceptible to misinfo, but misinfo impacts everyone.	people with low media literacy more susceptible to misinfo	
	older people more susceptible to misinformation	
	different media/presentations affect different groups	younger and older people are influenced by different media/presentations in terms of believability
	misinfo can target/affect everyone	
	media literacy as the goal	impossible to fact-check everything so improving the public's media literacy is the goal, but it could be a challenge
	multiple approaches to improve media literacy	comprehensive approaches such as educating people with different tools and about psychology theories; pre-bunking; learning from existing patterns of misinfo; and incorporating this training into educators, journalists, and policymakers' work.
Misinfo is closely related to current events local context	misinfo harm dependent on cultural/social context	

Continued on next page

(Continued)

	misinfo urgency related to social context	
	threshold of virality varies by countries	
	background knowledge essential to fact-checking	background knowledge includes local context, academics knowledge required for fact-checking
	misinfo induces/exacerbates hate	induce hate toward a person or group; stoke communal tensions
	vaccine hesitancy varies by country	
	distrust in authority exacerbates misinfo harm	define authority
	geographical divisions induce social conflict	
	relevancy / current events contribute to misinfo virality	
	triaging misinfo based on current events	relevance of the misinfo to current events and news cycle
	misinfo is generated by real-world events	
	triaging misinfo based on public interest	public interest is the welfare or well-being of the general public and society
Miscellaneous	convincing people by making fact-checking process transparent	
	publishing fact-checks to stop the spread of misinfo	publishing critical fact-checks is time-pressured. fact-checking to prevent harmful effects.
	difficulties in delivering fact-checks to people in need	
	context-checking to address opinions	

Continued on next page

(Continued)

	making fact-checks understandable by lay people	

C APPENDIX C: CODEBOOK FOR TOOLS

This section contains the interview codes related to automation tools that our interview participants suggested to be beneficial for fact-checking and content moderation. When it was necessary for clarifying among our coders, additional notes are included for more clearly defining the code.

Code	Notes on the Definition (Optional)
tool flagging content by potential to go viral	
tool showing the number of viewers	aggregated number of viewers across all platforms
tool detecting the recurrence of the same claims across media formats	
tool showing speed of being shared	
tool showing global view of misinformation spread	showing spread by country
tool accounting for popularity of the sender	
tool targeting the use of specific words	tool looking for buzzwords in claims
tool showing reach shares by demographics	
tool accounting for context	
tool comparable with local expert	
tool flagging content by sentiment level	
tool separating fact-checkable vs. noise	
tool detecting repeated claims	
automate debunking of repeated claims	e.g., matching claims with database of debunked claims
tool showing believability of a claim	
tool showing groups targeted by misinfo topics	
tool showing the misinfo's stage in life cycle	
tool converting videos to texts	to help determine what the video is trying to convey without having to watch them. exclude manipulated video.
tool facilitating transparent team decision-making	tool facilitating democratic decision-making
tool determining urgency by topic	
tool determining urgency by country priority	tool sensitive to information needs of a country
tools analyzing why something is harmful	
tool being language inclusive	