Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The double-edged sword of Twitter during crisis situations. A value-sensitive design approach to decreasing the impact of online misinformation

Yves van Engelen
TU Delft
Lavinia Marin (lavinia@xindi.ro)
TU Delft https://orcid.org/0000-0002-8283-947X

Research Article

Keywords: Misinformation, platform design, Twitter, crises, VSD, disaster response

Posted Date: August 3rd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1900328/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

Social media platforms like Twitter are extensively used during crises which seems to be a double-edged sword. On one hand, these platforms have an invaluable potential to become a reliable source of information and communication-tool for those involved in crisis situations. On the other hand, it is already known that current social media platforms allow misinformation to spread quickly - a fact that, in crisis situations, could widely jeopardize public safety. The reasons why misinformation spreads so quickly on Twitter in crisis situations can be traced back to several factors such as the platform's design choices, as well as the user's lack of critical engagement with the information, but, as we argue in this article, also because of an inherent value hierarchy embedded in the design of the user interface.

In this paper, we aim to first clarify what can be done effectively by Twitter to mitigate the spread of misinformation in crisis situations. By using a Value Sensitive Design framework, we argue that the current Twitter design promotes a multiplicity of values that are implemented in the user-interface design in ways that lead to conflicting effects on misinformation. We thus argue that a design intervention needs to change both the current value hierarchy and the design implementations of the values at stake. Our paper concludes with several design recommendations for elevating the user's critical engagement as well as a discussion about the moral implications of trying to streamline Twitter's user interface design towards favouring truth-telling in situations of crisis.

1. Introduction

In the fall of 2012, Hurricane Sandy hit Jamaica, the Dominican Republic, Haiti, Puerto Rico, Cuba, and the east coast of the USA (Gibbens, 2019, par. 5–15). Estimated is that Hurricane Sandy caused '$70 billion in damages' including flooding, mudslides, destructive winds, and many left homeless (Gibbens, 2019, par. 2–3). During this time roughly 20 million Tweets were sent on Twitter (Simon, Goldberg, & Adini, 2015, p. 616). Twitter was highly used by people to seek information updates about the storm (Gupta et al., 2013b, p. 729). Some users close to the disaster area tend to use the information on social media also as a base for the decision to evacuate or not (Sadri et al. in Wang & Zhuang, 2018, p. 1146). However, unfortunately misinformation was virally spread during this natural disaster using the same social media platform. A notorious example of misinformation was a post saying the floor of the New York Stock Exchange was flooded, with up to 3 feet of water (Hill; Holt; in Wang & Zhuang, 2018, p. 1148). This rumour was eventually spread by mainstream media (e.g. CNN) as well (Wemple in Wang & Zhuang, 2018, p. 1148). A more recent situation which emphasises the criticality of this research topic is the on-going Russia-Ukraine war where misinformation is spread widely as form of propaganda.

Mainstream social media platforms like Twitter are extensively used during crises, when there is high uncertainty and fast action needed to reduce damage or suffering from the event. On one hand, these platforms have an invaluable potential to become a reliable source of information and communication-tool for those involved in the crisis situation (Lee, 2019). People are increasingly using social media platforms to distribute and search for information related to ongoing live events (Starbird, Spiro, & Mason,

2015). This information is likely to be used by local communities and first responders to raise awareness, enhance decision-making, promote response efforts, and possibly save lives (Starbird et al., 2015). Major reasons why people rely upon social media platforms during a disaster are that these are convenient to use, have the ability to 'mass sending' and are known to be time and cost-effective (Peary, Shaw, and Takeuchi; White, Plotnick, Kushma, Hiltz, and Turoff in Abdullah et al., 2015, p. 32). On the other hand, it is already known that current social media platforms allow misinformation to spread quickly - a fact that, in crisis situations, could widely jeopardize public safety, public security, and emergency rescue attempts (Johnson, 2020; Abdullah et al., 2015). Misinformation, i.e., inaccurate or misleading information with or without the intention to cause harm, diffuses faster on social media than the truth (Vosoughi, Roy, & Aral, 2018). Social media platforms are fertile ground for misinformation because of the mass audience, the speedy exchange and spread of information, and because users seldom verify the information that they share (Tandoc, Lim, & Ling, 2017) while platforms "…lack the news media's editorial norms and processes for ensuring the accuracy and credibility of information" (Lazer et al., 2018, p. 1094). With roughly 330 million monthly active users (Twitter, 2019) and the characteristic of fast information spreading, the Twitter platform highly contributes to the distribution of misinformation worldwide albeit inadvertently.

Thus, the same platform characteristics that are valuable to use Twitter during crises are, at the same time, potentially dangerous for these users. This paper aims to untangle this conflict by looking at these platform characteristics through a Value Sensitive Design (VSD) framework, since we take it that this problem occurs in a complex social-technical environment where different values, incentives, and stakeholders are present. Although much research is done for tackling the problem of misinformation online in general, little research has been focused on ways of using design to improve user's critical engagement with the information consumed on a social media platform like Twitter, in particular during crises when people are more vulnerable to fall for misinformation. It is yet unknown how the problem of misinformation can be addressed by a design focussed approach. Thus, the main research question addressed in this paper concerns what kind of design-focused interventions can help platforms such as Twitter reduce the impact of misinformation sharing during crises?

To get a thorough answer, the next section will discuss why large-scale interventions against misinformation have not been taken yet by Twitter. Section 3 will present an empirical and conceptual analysis inspired by VSD framework in which we will clarify the values important for stakeholders, the values embedded in the platform design, what values should be promoted to combat misinformation, and what value conflicts are present. We aim to clarify what can be done effectively by Twitter to mitigate the spread of misinformation in crisis situations in section 4. Additionally, a discussion about the moral implications of trying to streamline the Twitter design towards truth-telling will be presented. Our paper concludes with several design recommendations for elevating the user's critical engagement in the final section.

2. The Current State Of Misinformation On Twitter

Misinformation is more present than ever on mainstream social media platforms like Twitter and this is partly due to the actions these platforms have or haven't taken in response to the waves of misinformation. Platforms are very aware of the presence and harmful effects of misinformation, as Twitter states "…we care deeply about the issue of misinformation and its potentially harmful effect on the civic and political discourse that is core to our mission" (Crowell, 2017, par. 1). However, current interventions by Twitter are ex-post; spreading misinformation and therefore violating Twitter's policy can lead to deletion of content, locking of the user's account, or even permanent suspension of the user (Twitter, n.d., par. 8). The policy is explicit, and the consequences are resolute, however misinformation is still present on the platform to a large extent (Zarocostas, 2020). It seems that these policies are by far not sufficient enough.

We hypothesise two main reasons why Twitter does not intervene ex-ante. Firstly, Twitter generates revenue by placing advertisements, for example, it generated up to '2617 million US dollars' in 2018 (Twitter, 2019). The longer time users spend on its platform, the more attractive Twitter becomes for advertisers to place their ads. As people are attracted to sensational or emotional content like "… gossip, rumor, scandal [and] innuendo" (Burkhardt, 2017, p. 8) which is highly present in misinformation (Marin, 2020), the incentive for Twitter to combat the enormous amount of misinformation conflicts with their purpose as a business, namely creating as much revenue as possible. Secondly, Twitter stated that it wants to stay away from censorship: "we, as a company, should not be the arbiter of truth" (Crowell, 2017, par. 3). Twitter depends on the self-correcting mechanism of Twitter users (including journalists, experts, and engaged citizens) (Crowell, 2017, par. 3) who will flag misinformation as they find it, yet the platform lacks quality control of the content created, shared, and consumed. Thus, Twitter emphasises it is being a platform and not a publisher in order to avoid legal constraints (Dvorak, 2018) and possible criticism of reducing freedom of expression.

Nonetheless, due to public pressure in recent years, Twitter needed to undertake some action in order to maintain its users. Recently Twitter implemented several minor (temporary) changes during the weeks before the 2020 US elections and performed its first public fact-check on several tweets of former president Trump (Gadde & Beykpour, 2020). This is an interesting tipping point, after years of not doing much, showing that Twitter is willing to take on a higher amount of responsibility, thus allowing for more initiatives to arise. However, for platforms like Twitter that want to be proactive about the misinformation shared on their site, what ex-ante measures are available? We turn to value sensitive design in order to sketch an answer to this question.

Value Sensitive Design Framework

Value sensitive design (VSD) as a framework is a "…theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process" (Friedman, Kahn, Borning, & Huldtgren, 2013, p. 56). This framework usually consists of three different investigations: conceptual, empirical, and technical (Friedman et al. 2013, pp. 57–58) which are re-enforcing each other. However, the methods of VSD may not be up to the problem we are

describing here: in crises situations, misinformation related to the crisis spreads rapidly because the sources cannot be verified easily (one needs to be there, on the ground, to check the hearsay and this is difficult even for professional journalists). There are several ways of designing a platform towards countering misinformation, however these ways are slow and rely on checking the new tweets against a curated database of known misinformation, and using human decision-makers to block or make invisible the misinformation tweets. In the case of an ongoing crisis situation, such checks are very hard to perform since there is no established truth and because the events are still unfolding. This makes the VSD analysis particularly tricky, especially for the conceptual analysis. In the conceptual phase of VSD, the researchers conduct empirical research into the values of the stakeholders (both direct and indirect) and try to understand what their value hierarchies are, as to be able to later implement these in the technical phase. Given the unpredictable nature of the crisis situations, it is very hard to gather enough stakeholders that might be affected by misinformation on Twitter (and the very target of the misinformation is hard to predict) to get their opinion on what could happen. This makes it that the conceptual phase of VSD in our case is purely speculative. The second tricky aspect is that we have identified two modes of using Twitter that are divergent: day to day use and crisis use. The same users that went on Twitter to be entertained may find themselves in situations where they are in the epicentre of a crisis situation and will have other needs from Twitter. These two uses make it very hard to establish a value hierarchy for Twitter right from the start. The main contribution of our paper consists in trying to solve this dual use and the value conflict it entails by proposing a design-based solution we will name *'a design switch*'. The second part of the VSD analysis focuses on the empirical phase of the VSD framework, in which we investigate which values are *actually* embedded in the current platform design. This empirical research was limited in scope but can be used as a blueprint for a larger scale VSD investigation of Twitter or other social media platforms.

We have conducted three interviews with main stakeholders and administered a survey amongst 12 Twitter users to identify values and user incentives in-depth in the empirical phase of the VSD methodology. The survey in the empirical phase was used to explore general insights of Twitter users.

3. The Values Of Twitter For Different Stakeholders

Values of different user groups

During crisis situations, when there is high uncertainty and limited information present, social media platforms are mainly used to share information and to seek information about the event. Social media platforms bring together different types of users. The first category of users is those directly affected by the crisis, in or close to the physical area. For an information-seeker, social media platforms tend to be a quicker way to gain knowledge about an event compared to mainstream media outlets (Oh et al. in Wang & Zhuang, 2018, pp. 1145–1146). However, accurate information and credible sources are difficult to find online. Users tend to find it difficult to assess the credibility of content shared online since, during a disaster, the biggest stream of content is from unknown sources, accounts with a smaller number of followers (Gupta et al., 2013b, p. 3). Additionally, fake content creators tend to cause confusion by

creating new accounts that are very similar to official account names (Gupta et al., 2013a, p. 9). Morris et al. (in Gupta et al., 2013b, p. 731) found that these 'visible' features like the username and the profile picture of the account were a key component for people to assess the credibility of the information. Hence, this can create confusion among users seeking information from 'trusted' organizations. Furthermore, during crises and emergencies, the assessment of information is also more difficult as "due to heightened anxiety and emotional vulnerabilities, people are often more susceptible to fall for rumors / fake content" (Gupta et al., 2013a, p. 1). Rassin (2008) states that it is difficult for people to change their belief about a topic, even if there is supported evidence that is in conflict with this belief. Due to this confirmation bias, people prefer information that is in line with their already existing attitude (Kim, Moravec, & Dennis, 2017, p. 933). Additionally, they tend to "ignore information that challenges them" (Kim, Moravec, & Dennis, 2017, p. 933). To conclude, the assessment of whether the information is true is done very quickly by users while verification is too much of a hassle, resulting in less accurate information spread overall. Even more, this assessment of information is jeopardized by the confirmation bias of the information consumers, who often rely on cognitive heuristics to decide what is true (Koroleva et al., 2010, p. 5).

The other group is constituted by those users not directly affected by the crisis, those less dependent on the information shared. Commonly, these users are less informed on what is going on and tend to share content more frequently without thinking about it twice. As Twitter brings together different types of content for different purposes, both from official news agencies but mostly generated by regular users, a blur of boundaries between news and other types of content can limit the critical engagement people have with the news they consume (Ofcom, 2018). For example, when one is scrolling through its feed for entertainment purposes and sees a post on an earthquake happening at the other side of the world, one could share this to spread awareness or as an approach to 'help' others, without thinking about the consequences it might have in case this was not true. Users of Twitter only take a very short time to evaluate the credibility of a Tweet on which they base subsequent action on (to share, ignore, like, etc.) (Wierzbicki, 2018, p. 136). Most people tend to share articles on social media after only having read the headline (Gabielkov, Ramachandran, Chaintreau, & Legout in Kim, Moravec, & Dennis, 2017, p. 939). Tandoc, Lim, & Ling (2017) found that "users seldom verify the information that they share" (p. 139). Gupta et al. (2013a) found that content that went viral was forwarded by a lot of verified accounts as well (with a large number of followers), causing misinformation to have an even bigger reach. Wang & Zhuang (2018) found that "not all social media users will combat rumours even if they have already been debunked by accurate information" (p. 1147). Achieving more clarity by the 'self-correcting mechanism' of users described by Twitter to combat misinformation is done minimally since events happen too fast in crisis situations, followed by user reports, and the checking mechanism is overwhelmed by the amount of information coming from users.

The different incentives of these two user groups to use the platform show a conflict of values prioritised. Those directly affected by the crisis prioritize accuracy of information and credibility of information sources to gain more knowledge and certainty. Those not directly affected by the crisis value knowledge for entertainment and a good user experience of the platform. Based on the findings above summarised

from the literature on misinformation in crisis situations on social media, a list of values of both user groups is displayed in Table 1.

Table 1
–Values related to misinformation

Values of directly affected stakeholders	Values of indirect stakeholders
Accuracy of information	Entertainment / amusement
Credibility (of user and information)	Usability
Knowledge	Knowledge
Certainty	Rapidity of interaction

Values embedded in the platform design

The ensuing question is then which of the values above are embedded in the design of the Twitter platform. Van de Poel (2013) argues that values form the basis for formulating norms of behaviour, which eventually can be translated into design requirements of the technology. The design of Twitter incorporates the platform's values, the developer's, but also presumably some values of their users. However, what values emerge as most obvious for the users interacting with the Twitter interface in a crisis situation? We turned to an empirical investigation for this answer: for this, we will first analyse those design features that make Twitter a heaven for misinformation in crisis situations and then analyse the values embedded in these design choices.

The Twitter platform has an open character, allowing anyone to register for as many accounts as they want, anonymously, and free. This open character is a strength, as it contributes to freedom of speech and accessibility for those affected on-site of the crisis. However, at the same time, this creates a big drawback; as (most) users' identity is not verified, they can create accounts under anonymous names, and then can share as much misinformation as they want. At the same time, all users are stimulated to engage with the information they consume and create content themselves. While going through the content ('feed'), the icon to create content ('tweet') is glaringly visible to the user at any time. Additionally, users can comment on, share ('retweet'), and like the content they see, all with just one click. Twitter makes it effortless to engage with information, users do not even need to take the time to assess what they are actually engaging with. Specifically, the 'retweet' button enables the propagation of misinformation, as the information "loses connection to its original author, time, and the context in which it was shared" (Starbird et al., 2014, p. 655). This makes it harder for users to assess the credibility of the content. A design choice unique to Twitter is the limited number of characters per post. This character-limit drives users to add hyperlinks to other websites for additional information, which enables the spread of misinformation as "the link sharer can avoid responsibility" (Shin, Jian, Driscoll, & Bar, 2018, p. 280–281).

Currently, Twitter uses algorithms to optimize user's engagement. The content shown to Twitter users first, on the top of the page, will be the content that Twitter algorithms predict that the user will like most. This might be beneficial for users at first, as they consume more content they like. However, this algorithm leads to the so-called 'filter bubbles' or 'echo chambers' (Pariser, 2011). As the algorithm automatically filters out information that the user might not engage with that much, it actually leads to a decrease in the diversity of content users consume (Bozdag & van den Hoven, 2015, p. 249). These filter bubbles are detrimental because of the tendency to lead to confirmation bias; users tend to believe information that favours their pre-existing beliefs, even if this information might be inaccurate or false. Important to note is that people often do not realize they are inside a filter bubble and will not know what they are missing out on (Pariser in Bozdag & van den Hoven, 2015, p. 249).

The next factor enabling the spread of misinformation through the design of the platform is the presence of 'bots'. Bots are automated, software-driven "accounts that participate in news and information dissemination on social networks" (Lokot & Diakopoulos, 2015, p. 682). An estimation by Varol et al. (2017) suggests that up to "15% of active Twitter accounts are bots" (p.1). Bots are also increasingly used for financial gains (Weller et al. 2014, p. 185) and for sharing malicious content. The capacity of a bot to spread fake content is enormous compared to what a single person is able to. Bots are not only used to spread content but also to create fake 'followers'. Bot 'followers' can be bought on the black market (Weller et al., 2014, pp. 185). This is important, because, as previously stated, the number of followers is likely to contribute to the credibility of an account (K. P. K. Kumar & Geethakumari, 2014). This results in the ability to purchase some sort of credibility, by buying followers that are operated automatically. Consequently, the presence of algorithms for engagement and 'bots' on Twitter contributes to the propagation of misinformation.

The absence of a large-scale quality control system is also contributing to the propagation of misinformation. Social media platforms differ from traditional media companies in the way they ensure the accuracy of information. There are no "editorial norms and processes" present (Lazer et al, 2018, p. 1094). Specialized skills or training required at traditional media companies are not needed (Agarwal, 2010, p. 4). Publishers (content creators) are driven by speed over quality. In our digital age, speed is getting more important for publishers, as 'everyone' wants to be the first to write about breaking events. This results in the lack of quality of some content, as the author has little time to verify if the content is fully accurate. One of the possible solutions to ensure the quality of content is fact-checking. However, fact-checking is not implemented (yet) on the Twitter platform. Fact-checking can be done manually or automatically. One of the drawbacks of individual, manual fact-checkers is that it is done long after the content has already been created and shared (Kim & Dennis, 2018, p. 3956), making it useless in case of crises. Automated fact-checkers rely on the "… availability of verifiable sources…", which is complicated and a typical human task (Bell, 2019, par. 10) and, in the particular case of crises, not sufficient. Fact-checking is therefore not sufficient and scalable during crises. To summarize, a high level of accessibility and the culture of 'publishing first' and checking later encourages that most users will value speed over accuracy, especially those users that strive to become influencers.

The take-away point from this section is that the current design features of Twitter promote values that contribute to the phenomenon of misinformation. At the same time, values that should be promoted in order to reduce the impact of misinformation in crisis situations are downplayed. Table 2 displays an overview of the values promoted and downplayed by the current design features.

Table 2
– Values embedded in the platform design

Values of Twitter translated in design features	Values of Twitter downplayed by design
Accessibility	Responsibility
Speed	Credibility
Confirmation	Clarity
Amusement	Accuracy
Engagement	Knowledge

In conclusion, Twitter promotes a multiplicity of values that are implemented in the user-interface design in ways that lead to disruptive effects on information shared by users, especially in crisis situations. We have identified four main value-conflicts between the different user groups identified in the previous section.

• first, between the value of entertainment and knowledge. The creation of disinformation is driven by the desire for sensational and emotional content, something encouraged by the current design features, e.g., the engagement algorithm, infinite scrolling, and the visibility of engagement indicators. The value of entertainment for users not directly affected by a crisis sabotages the value of knowledge for those directly affected by the crisis.

• second, between speed and accuracy. Accuracy of information is valued by those directly affected as it can increase safety during crises. However, information is often shared before verified by those not directly affected by the crisis. This reveals that the value of speed sabotages the value of accuracy. Here as well, the design features neglect to promote accuracy and at the same time discourage speed of sharing, e.g., the ‘one-touch’ share or like button.

• third, between the desire for certainty and the need for clarity within the group of directly affected people. Just as accuracy, clarity is valued as it can increase the user’s safety. However, people will make up their ‘own versions of the truth’ when there is high uncertainty because they want to believe certain states of affairs more than others. This desire takes place in our everyday lives and has been studied under the name of confirmation bias – namely the tendency to see out there only one confirms one’s pre-existing beliefs. Through the personalisation algorithms, Twitter (just as many other mainstream social media platforms) shows users what they want to see. However, in a crisis situation, showing the users what they want to see, especially if they are directly affected by the disaster, will undermine the value of certainty and hence their possibilities for action.

• The fourth value conflict is not per se between different user groups, but between short- and long-term values, between accessibility and credibility. Accessibility is valued on the short-term as it enables people to gain quick information. However, this accessibility can sabotage the long-term value of credibility as anyone can be an information provider.

Should there be a design-switch for the Twitter interface during emergencies?

When users access their phones at night, certain apps display a dark background – usually to make it easier on the eyes of the users; however, this dark background has the additional and unintended effect that it signals to the user that it is dark outside, hence late. Some people do not raise their heads out of their phones enough to look out the window, but the background switch on their phone informs them of the sunset. We have imagined a similar signalling function on Twitter for crisis situations and we have called it the 'design switch'. Here is how it functions: when a user enjoys Twitter on a day to day basis, nothing looks different from the current design. But once the user is in a crisis zone or is about to retweet about a current crisis, the design of Twitter will change visibly, with several backgrounds and functionality changes, as to signal to the user that they are entering a 'crisis' or 'disaster' mode. In this mode, the truthfulness and accuracy of Tweets become paramount and other values are downplayed by design. The design recommendations and the norms we propose in the next sections are meant to function as a 'design switch' only in situations of emergency such as in the case of terrorist attacks, natural disasters, wars, civil revolutions, or riots. For day-to-day operations, Twitter can revert to its usual design of the user interface that promotes engagement, entertainment, speed of information, and even allows for confirmation-bias. The ethical implications for our recommendations would then revolve around the necessity for this design switch to take place at all and for the necessary conditions that would trigger such a design switch. In this section, we want to argue for the necessity of the design-switch and for its being an advisable design intervention.

In order to justify if such a design switch should happen, we need to compare it with the alternative options: either Twitter does nothing and keeps its current design even in crises situations, or some other design intervention could be proposed. The 'do nothing' alternative seems morally problematic since Twitter is responsible at least partially for how people use it during crisis situations. Twitter puts at people's disposal a powerful tool for making their voices heard with virtually no accountability. Misinformation sharing is akin to engaging in rumour or gossiping (Marin, 2021). During everyday activities, gossip and rumour do not hurt that much; but when the stakes are high – and this is the case with crises – then suddenly this gossip online receives unwarranted attention. People follow the hashtags related to the crisis and come across pieces of gossip that would have never gotten that much visibility. This inflated visibility would not have been possible with offline situations of gossip, and Twitter makes it possible through its infrastructure of hashtags and through making the trending hashtags visible to all users. Because it warrants visibility without accountability, we think that Twitter needs to assume at least some part of the responsibility for the visibility of misinformation during crises. Hence the option to do nothing cannot hold. The other option, as an alternative to the design switch, would be to ask Twitter to implement a pre-check on accuracy all the time, in its day-to-day operations[1]. But this would mean

imposing a value hierarchy on Twitter that Twitter itself and its users might reject. After all, many users come for entertainment on this platform and to ask them, in non-emergency situations, to check their Tweets as a journalist does would potentially make users want to leave this platform. This is why we do not think it realistic to change the day-to-day interface of Twitter towards accuracy and truth-telling.

A possible counter-argument to the design switch could be that we are proposing the limiting of the freedom of speech for regular users in emergency situations. However, in cases of social emergencies, certain day-to-day freedoms or practices need to be restricted because in those situations the rights of the vulnerable few (those affected by the crisis) take precedence over the rights of the unaffected many (such as the right to be informed or entertained for the bystanders). As seen in the case of the COVID-19 pandemic, when certain individual freedoms to circulate, to go out at night, to gather with friends for parties were restricted – the justificatory principle used was that one's individual freedom stops where the harm to others begins (Ferdinand, 2021). The idea here is to weigh the need for people to speak up their minds – even if this means using unverified information – versus the collective need that, in the life-threatening situation, lives are saved and people are calm enough to take the appropriate measures. As van den Hoven already argued, there is no freedom of speech when someone shouts "Fire!" in a crowded theatre-hall (van den Hoven, 2008).

It could also be argued that the value conflict at stake in Twitter-popularised emergencies is not primarily about accuracy versus entertainment. Many users who misinform others in disaster situations may not seek entertainment, but want to be genuinely helpful. Thus, their choice to click "retweet" on something they just saw on their news feed may be motivated by simply wanting to make the situation at stake rapidly known to others. However, even in these situations, the value of accuracy should trump that of speed because once the fire of misinformation is ignited, it is very hard to stop, and the damage done cannot be taken back by a retraction. In these cases, prudential principles apply because the amount of harm done by one tweet cannot be estimated realistically and it is better to err on the safe side.

The design switch proposed here assumes that people use Twitter in their day-to-day life primarily for entertainment and amusement purposes. Yet many people use Twitter as one of their most valuable sources of information, as they are following trustworthy users by actively curating their news feed. Would the design switch make sense for these users? We think that yes, since anyone can make mistakes in an emergency situation when the judgement may be clouded by the desire to help. In this respect, we do not think that there are more or less critical users, rather users who can keep their thinking under pressure and that always carefully check the information they pass on, and users who forget about this need because, for them, other concerns become paramount. The design switch is a proactive measure that is not about delivering to users the truth necessarily but about nudging them to be more careful in how they publicly speak about a crisis situation. We want to stress that the measures we proposed are not about treating the user as a passive actor, but that encourages the user to think critically about the information one is about to post or share. In this sense, we think that our design switch is better understood as a critical switch as it is nudging users to use their own faculties for critical thinking when the situation at stake is important enough.

[1] This possibility is not that far-fetched, as currently (at the time of writing this article) the European Commission is working at the Digital Services Act in which social media companies will be held directly responsible for the misinformation and hate-speech trafficked on their platforms.
See: https://oeil.secure.europarl.europa.eu/oeil/popups/ficheprocedure.do?reference=2020/0361(COD)&l=en

4. Design Recommendations

In this section, we aim to clarify what can be done by design effectively by Twitter to mitigate the spread of misinformation in crisis situations. The presence of misinformation in such situations threatens the values of accuracy, clarity, credibility, and knowledge. In addition, the absence of the feeling of responsibility to combat misinformation from the user base contributes to the great impact of misinformation. To decrease the impact of misinformation during a crisis, users need to be challenged and encouraged to take responsibility. Additionally, the link to prioritising the values of entertainment, certainty, and speed need to be discouraged during a disaster, to reduce the creation and propagation of misinformation. Accessibility is part of a value conflict as well, however, accessibility is very beneficial during crisis-situations. Therefore, accessibility should not be discouraged. Additionally, the value of the freedom of expression contributes to the creation of misinformation; however, as discussed in the previous section, it is not clear whether this value should be sacrificed in order to reduce the creation of misinformation. Concerning the feasibility of our proposal, we have an example that it can work: during the 2020 US elections, Twitter proved to be able and willing to make such temporary changes. The platform temporarily changed how the retweet-feature worked, encouraging users to comment on the tweet before sharing it (Peters, 2020).

The value hierarchy as it is embedded in the current Twitter user-interface design seems insufficient to reduce the impact of misinformation during crises. During crises, a (temporary) change is needed to embed the priority values within the design of the user interface. Design changes should be oriented towards the promotion of truth-telling in situations of crisis. As the impact and reach of Twitter and therefore the misinformation spread is so big, the current design features to increase user engagement and therefore revenue should come to the second place if saving lives can be achieved. Several examples of design principles that can help achieve this are listed below:

Firstly, the focus of design should lay on increasing the quality and consciousness of the actions of the users.

The way users interact with content on Twitter needs to change in a way that they are more aware of the problem, more considerate when they share something, and encouraged to be more conscious and critical.

Secondly, the focus of design should lay on promoting the values of directly affected people and

discouraging the values of not-directly affected people. By explicitly promoting the values accuracy, certainty, credibility, and knowledge into the design features of Twitter and discouraging the short-term values entertainment, user engagement, and rapidity of interaction in case of an emergency, the impact of misinformation can be reduced.

Thirdly, design features should encourage users to have a more active role in combating misinformation.

The problem of misinformation can only be effectively encountered by a change of behaviour among users; one where users take responsibility and an active approach to debunk misinformation.

It is recommended to investigate how the enthusiasm and willingness of those not-directly affected to help people during a disaster can be implemented in a way that is beneficial for emergency services. At the moment, this enthusiasm and attempt to help is not beneficial for the safety of people and only creates 'irrelevant noise'.

In VSD, the choice of design features is based on first translating values into norms and then norms into design features (Van de Poel, 2013). Based on the three norms or design principles that we proposed, we find that several possible design features, as displayed in Table 3, could contribute to slowing down the propagation of misinformation on Twitter during disasters. Three of these design features will be explained in more detail below.

Table 3
–Proposed design features

Design features
A green verified button to distinguish official emergency-related accounts
An exclusive page for aggregating verified well-trusted sources
Reliability ranking
Delay to retweet
Making it easier for people to report inaccurate information
A notification to people that shared or liked the Tweet that has been debunked
Indication of trustworthiness by other users

Delay to retweet

To decrease the number of users sharing articles while only have read the headline and not in its entirety and to promote verifying information with other sources, a delay to share content can be added. The identified pay-off between accuracy and speed during emergencies should be rearranged. As users have less time to verify, the accuracy of the information shared decreases. This proposed feature, as illustrated in Fig. 1, will give a notification with the question 'This Tweet is not yet confirmed by other sources. Are

you sure you want to retweet this Tweet?'. This pop-up will not be present on every Tweet, only those related to the disaster and not confirmed yet. The other feature in Fig. 1 will give a notification with the question 'We noticed you have not read the article yet. Are you sure you want to retweet this Tweet?'. This pop-up will not be present on every Tweet, only those when the user did not ‘read’ the article shared yet (based on the time it takes someone to read an article versus to read the title). Twitter announced a similar test-feature during the US 2020 elections. Twitter’s feature would only show a delay when a user wants to retweet a Tweet containing a link to an article the user has not yet read on Twitter (Porterfield, 2020). Additionally, users were encouraged to add a comment to the tweet they were about to share.

The idea behind this 'delay' is to encourage users to give it a second thought and be more critical of the information before directly sharing it. In the end, a more critical mindset will lead to a smaller impact of misinformation. This feature contributes to the awareness and responsibility of users and discourages the rapidity of sharing. The effectiveness of the implementation of this feature during the US 2020 elections is not studied yet, but we suggest using this temporary change during crisis situations as well.

An exclusive page of verified well-trusted sources

Findings show that one of the motives to use Twitter during a crisis is to seek information because traditional mass-media channels are not yet reporting on the situation, or not covering enough the event. This absence of information causes people to make up their ‘own versions of the truth’, leading to rumours and misinformation. As a result of a confirmation bias, users tend to believe information that favours their worldview, even if this information might be inaccurate or false. To counteract this tendency, Twitter needs to give users the possibility to find accurate information all in one place, thus discouraging the tendency for confirmation bias which arises when users can pick and choose their preferred sources.

The proposed feature will show a red icon of an exclamation mark at the top of the page when an emergency-event is happening, as seen in Fig. 2. Users have the ability to click on this icon and see a message that allows them to switch to an alternate feed, exclusive for content about the ongoing event (e.g., an earthquake, a shooting, a hurricane). The content on this page is only coming from recognized and relevant sources such as local police stations, the government, or rescue teams.

The idea behind this feature is to enable users to get trustworthy, local information in situations of high uncertainty. Accurate information can enable users to get out of their filter bubble and think more critically by seeing information that is conflicting with their pre-existing attitude. If users are not open to the idea of getting different content that might be conflicting with the sort of content they usually consume, the feature will not enable people to get out of their filter bubble. Therefore, this feature will not give the control to the user but will automatically filter-in the information from this feed. Users can see posts in this feed from accounts they normally do not follow but are nevertheless relevant for the ongoing disaster. A point of contestation could be that the choice of which accounts are shown, and which accounts are left out is done by Twitter. However, in case of emergencies safety is the priority and the competition between accounts should not be too important.

A (temporarily) green verified button to distinguish official emergency-related accounts

It is hard for users to distinguish trustworthy sources from fake sources. A critical factor found is that the reach of misinformation is so big because verified accounts spread misinformation as well (Gupta et al., 2013a). These verified accounts are trusted by users and most of the time have a large following base. Users have gained familiarity and a build-up trust with verified accounts during normal circumstances. However, during crisis-situations users cannot fully trust these verified accounts making it even harder to assess the content as being true or false. The ability to assess the credibility of information is reduced during crises, as people are more emotionally vulnerable to fall for inaccurate information. Emotions during crises (e.g., fear, anxiety, sadness) tend to influence the perception of people by assessing the truthfulness of the content.

The proposed feature will give relevant accounts a green verified button instead of a blue verified button, to make a distinction between 'normal' verified accounts and verified accounts that are relevant during crises. A visual example of the feature is shown in Fig. 3. The green verified icon will be given out to official organisations relevant during emergencies, e.g., the local police station, rescue teams, national and local governments, etc. The green verified icon will only be activated during the event. The goal of this feature is to help users by identifying trustworthy sources which are people on the ground, rescue teams, thus those users who suddenly become visible because of their involvement in the disaster but which one would not normally follow.

User testing of the proposed design features – first impressions

The three features discussed above have been shown to 12 Twitter users by using a survey to get their first impression on the feasibility and effectiveness.

The first impressions on the *Delay to (re)tweet* (Fig. 1) feature were diverse, both positive and negative. According to the users, this feature represented the values awareness, knowledge, and safety. Some respondents see the value of this feature and support the effectiveness of it: "*This is a remarkable one, on the first side I think this is a feature that makes sure you know what you're reposting. It makes you aware of what you're doing*!" and "*Good idea, it will indeed probably lead to fewer retweets of unverified articles and Tweets*". However, another respondent is more critical about the effectiveness of the feature: "*Not sure if it has the necessary impact. Two clicks is not much more than one*". This is a relevant point and therefore this feature might not be relevant for all users. Nevertheless, if it will affect some users, it can already help to reduce the impact of misinformation propagation.

The first impressions on the feature of the *Alternate feed with trustworthy sources* (Fig. 2) were mostly positive and according to the users represented the values of safety, security, knowledge, and accuracy. Some respondents were very excited about this feature and pointed out that it can increase safety and accuracy of information: "*This could be lifesaving"* and "*This is exactly what Twitter should have, for people in an emergency situation it's sometimes difficult to find the 'right' information from the 'right' sources*". Points of criticism were focussed on what accounts will be displayed *"… what sources to*

display, and what not?" and the effectiveness: "*often you see that smaller accounts have the insight information (also fake sometimes) so people still will search beside big organizations*". We argue that this decision differs depending on the crisis that is going on and can be made by best practices once the feature is implemented.

The first impressions on the *Green verified icon for local trustworthy sources* (Fig. 3) were diverse as well. Some positive comments were: "*A great feature, the Twitter users are directly informed about the trustworthiness*" and "*This seems very promising, I think if communicated right to the users this could really work*". Some more critical comments pointed out that this feature needs to be explained in the right way. As indicated from the comments below, users might think that organisations can 'choose' what colour of the verified button they want: "*[I] think it's really helpful, but every news profile or popular figure would want this button. I think that in practice everyone would shift from blue to green (or a big group)*". Another respondent also stated that it is not totally clear how the process works: "*To me, the elaboration is not yet clear: who is going to decide who gets a green verified button? …, How is the trustworthiness guaranteed?*". Therefore, the communication on this feature needs to be transparent and explicit so users will exactly know why a certain account has a green verified button.

5. Conclusions

The double-edged sword of social media platforms like Twitter should become a one-edged sword in case of crises situations. To ensure the invaluable potential of being a reliable source of information and communication-tool for those involved in the crisis situation, the impact of the spread of misinformation needs to be reduced. This research shows that the reasons why misinformation spreads so quickly on Twitter in crisis situations can be traced back to the design choices of the platform, as well as the user's lack of critical engagement with the information, but, most crucially, also because of the value hierarchy embedded in the user interface design. We argue that the current Twitter design promotes a multiplicity of values through the user-interface design in ways that lead to conflicting effects on misinformation. Design interventions aimed at tackling misinformation on Twitter need to change both the current value hierarchy and the design implementation of the values at stake. Several design changes have been recommended by using multiple design principles and we have tested the feasibility of these design features with a limited number of users.

The necessity of social media platforms to take action into combating misinformation has increased remarkably with the impact it has during crises situations. As Twitter already showed the willingness to take on a higher amount of responsibility (e.g., fact-checking during the US 2020 elections), the recommended design changes have high feasibility of being implemented. The proposed design changes are not in conflict with values as freedom of expression, nor are they in conflict with the usability and revenue model of Twitter as they are targeted to help the information consumer. The design changes enable a small reform with possibly big implications for those affected by a crisis. As the problem of misinformation is comprised of different elements, so is combating it. Albeit design changes are only a part of the puzzle, they can still have a great effect on reducing the impact of misinformation. We

recommend further research to investigate the long-term effectiveness of implementing such design changes on Twitter and other social media platforms and how user-platform interactions can be shaped to optimize the benefits during crises.

Declarations

Competing interests: The authors declare no competing interests.

Participant consent: All participants interviewed were informed verbally, before the interview, of their right to withdraw at any time and that their contributions will be kept fully anonymous. All participants consented verbally to take part in the interviews. All participants in the online survey read an informed consent notice and agreed to it before proceeding with the online survey. There was no ethics approval needed from the university since the study was part of a Bachelor's thesis which is exempt from this requirement.

References

Abdullah, N. A., Nishioka, D., Tanaka, Y., & Murayama, Y. (2015). Action and Decision Making of Retweet Messages towards Reducing Misinformation Spread during Disaster. *Journal of Information Processing*, *23*(1), 31–40. https://doi.org/10.2197/ipsjjip.23.31

Agarwal, N. (2010). Information quality challenges in social media. *Department of Information Science The University of Arkansas at Little Rock*, 1–16. Retrieved from https://www.researchgate.net/publication/260337476

Bell, E. (2019). The Fact-Check Industry. Retrieved 17 May 2020, from https://www.cjr.org/special_report/fact-check-industry-Twitter.php

Bozdag, E., & van den Hoven, J. (2015). Breaking the filter bubble: democracy and design. *Ethics and Information Technology*, *17*(4), 249–265. https://doi.org/10.1007/s10676-015-9380-y

Burkhardt, J. M. (2017). History of Fake News. *Combatting Fake News in the Digital Age Chapter 1 of Library Technology Reports*, *53*(8), 5–9. Retrieved from https://journals.ala.org/index.php/ltr/article/view/6497

Crowell, C. (2017, June 14). Our approach to bots and misinformation. Retrieved 10 May 2020, from https://blog.Twitter.com/en_us/topics/company/2017/Our-Approach-Bots-Misinformation.html

Dvorak, J. C. (2018, August 15). Twitter and Facebook Are Publishers, Not Platforms. Retrieved 11 May 2020, from https://www.pcmag.com/opinions/Twitter-and-facebook-are-publishers-not-platforms

Ferdinand, K. C. (2021). COVID-19 Mitigation: Individual Freedom Should Not Impede Public Health. *American Journal of Public Health*, 111(4), 592-593.

Figueira, Á., & Oliveira, L. (2017). The current state of fake news: challenges and opportunities. *Procedia Computer Science*, *121*, 817–825. https://doi.org/10.1016/j.procs.2017.11.106

Friedman, B., Kahn, P. H., Jr., Borning, A., & Huldtgren, A. (2013). Value Sensitive Design and Information Systems. *Early Engagement and New Technologies: Opening up the Laboratory*, 55–95. https://doi.org/10.1007/978-94-007-7844-3_4

Gadde, V., & Beykpour, K. (2020, October 9). *Additional steps we're taking ahead of the 2020 US Election*. Twitter. https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes.html

Gibbens, S. (2019, February 11). Hurricane Sandy, explained. Retrieved 30 May 2020, from https://www.nationalgeographic.com/environment/natural-disasters/reference/hurricane-sandy/

Gupta, A., Lamba, H., & Kumaraguru, P. (2013a). $1.00 per RT #BostonMarathon #PrayForBoston: Analyzing fake content on Twitter. *2013 APWG ECrime Researchers Summit*. https://doi.org/10.1109/ecrs.2013.6805772

Gupta, A., Lamba, H., Kumaraguru, P., & Joshi, A. (2013b). Faking Sandy. *Proceedings of the 22nd International Conference on World Wide Web - WWW '13 Companion*. https://doi.org/10.1145/2487788.2488033

Huang, Y. L., Starbird, K., Orand, M., Stanek, S. A., & Pedersen, H. T. (2015). Connected Through Crisis. *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing - CSCW '15*, 969–980. https://doi.org/10.1145/2675133.2675202

Johnson, B. (2020, January 24). Fake News During Disasters Putting People's Lives at Risk, Warns Intel Bulletin - Homeland Security Today. Retrieved 1 March 2020, from https://www.hstoday.us/subject-matter-areas/emergency-preparedness/fake-news-during-disasters-putting-peoples-lives-at-risk-warns-intel-bulletin/

Kim, A., & Dennis, A. (2018). Says Who?: How News Presentation Format Influences Perceived Believability and the Engagement Level of Social Media Users. *Proceedings of the 51st Hawaii International Conference on System Sciences*, 3955–3965. https://doi.org/10.24251/hicss.2018.497

Kim, A., Moravec, P., & Dennis, A. R. (2017). Behind the Stars: The Effects of News Source Ratings on Fake News in Social Media. *SSRN Electronic Journal*, *36*(3), 931–968. https://doi.org/10.2139/ssrn.3090355

Koroleva, K., Krasnova, H., & Günther, O. (2010). Stop spamming me! *exploring information overload on Facebook. In: Proc 16th AMCIS, Lima*.

Kumar, K. P. K., & Geethakumari, G. (2014). Detecting misinformation in online social networks using cognitive psychology. *Human-Centric Computing and Information Sciences*, *4*(1), 1–22. https://doi.org/10.1186/s13673-014-0014-x

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., … Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lee, K. S. (2019). Explicit Disaster Response Features in Social Media. Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services - MobileHCI '19 https://doi.org/10.1145/3338286.3340140

Lokot, T., & Diakopoulos, N. (2015). News Bots. *Digital Journalism*, *4*(6), 682–699. https://doi.org/10.1080/21670811.2015.1081822

Marin, L. (2020). Three contextual dimensions of information on social media: lessons learned from the COVID-19 infodemic. *Ethics and Information Technology*. doi:10.1007/s10676-020-09550-2

Marin, L. (2021). Sharing (mis) information on social networking sites. An exploration of the norms for distributing content authored by others. *Ethics and Information Technology*, 1–10. doi:10.1007/s10676-021-09578-y

Ofcom. (2018). *Scrolling news: The changing face of online news consumption*. Retrieved from https://www.ofcom.org.uk/__data/assets/pdf_file/0022/115915/Scrolling-News.pdf

Pariser, E. (2011). *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin.

Peters, J. (2020, 21 oktober). *How to retweet using Twitter's new temporary format*. The Verge. https://www.theverge.com/21524092/twitter-temporarily-changing-retweet-quote-tweet-election

Porterfield, C. (2020, June 10). Twitter Begins Asking Users To Actually Read Articles Before Sharing Them.

Retrieved 21 June 2020, from https://www.forbes.com/sites/carlieporterfield/2020/06/10/Twitter-begins-asking-users-to-actually-read-articles-before-sharing-them/#1dd356fd66a3

Rassin, E. (2008). Individual differences in the susceptibility to confirmation bias. *Netherlands Journal of Psychology*, *64*(2), 87–93. https://doi.org/10.1007/bf03076410

Shin, J., Jian, L., Driscoll, K., & Bar, F. (2018). The diffusion of misinformation on social media: Temporal pattern, message, and source. *Computers in Human Behavior*, *83*, 278–287. https://doi.org/10.1016/j.chb.2018.02.008

Simon, T., Goldberg, A., & Adini, B. (2015). Socializing in emergencies—A review of the use of social media in emergency situations. International Journal of Information Management, 35(5), 609–619. https://doi.org/10.1016/j.ijinfomgt.2015.07.001

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing. *IConference 2014 Proceedings*, 654–662. https://doi.org/10.9776/14308

Starbird, K., Spiro, E., & Mason, R. (2015, September 30). Detecting Misinformation Flows in Social Media Spaces During Crisis Events. Retrieved 21 March 2020, from https://faculty.washington.edu/espiro/project/misinformation/

Tandoc, E. C., Jr., Lim, Z. W., & Ling, R. (2017). Defining “Fake News”. *Digital Journalism*, *6*(2), 137–153. https://doi.org/10.1080/21670811.2017.1360143

Twitter. (2019, April 23). Q1 2019 Earnings Report [Slides]. Retrieved from https://s22.q4cdn.com/826641620/files/doc_financials/2019/q1/Q1-2019-Slide-Presentation.pdf

Twitter. (n.d.). Platform manipulation. Retrieved 10 May 2020, from https://transparency.Twitter.com/en/platform-manipulation.html

Van Den Hoven, J. (2008). Information technology, privacy, and the protection of personal data. *Information technology and moral philosophy*, 301-321.

Van de Poel, I. (2013). Translating Values into Design Requirements. *Philosophy and Engineering: Reflections on Practice, Principles and Process*, 253–266. https://doi.org/10.1007/978-94-007-7762-0_20

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online Human-Bot Interactions: Detection, Estimation, and Characterization. *Social and Information Networks (Cs.SI)*, 1–10. Retrieved from https://arxiv.org/abs/1703.03107

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146– 1151. https://doi.org/10.1126/science.aap9559

Wang, B., & Zhuang, J. (2018). Rumor response, debunking response, and decision makings of misinformed Twitter users during disasters. *Natural Hazards*, *93*(3), 1145–1162. https://doi.org/10.1007/s11069-018-3344-6

Weller, K., Bruns, A., Burgess, J., Mahrt, M., & Puschmann, C. (2014). *Twitter and society*. Retrieved from https://nbn-resolving.org/urn:nbn:de:0168-ssoar-47764-2

Wierzbicki, A. (2018). *Web Content Credibility*. Retrieved from https://doi.org/10.1007/978-3-319-77794-8

Zarocostas, J. (2020). How to fight an infodemic. *The Lancet*, *395*(10225), 676. https://doi.org/10.1016/s0140-6736(20)30461-x

Figures

Figure 1

Delaying the retweet function

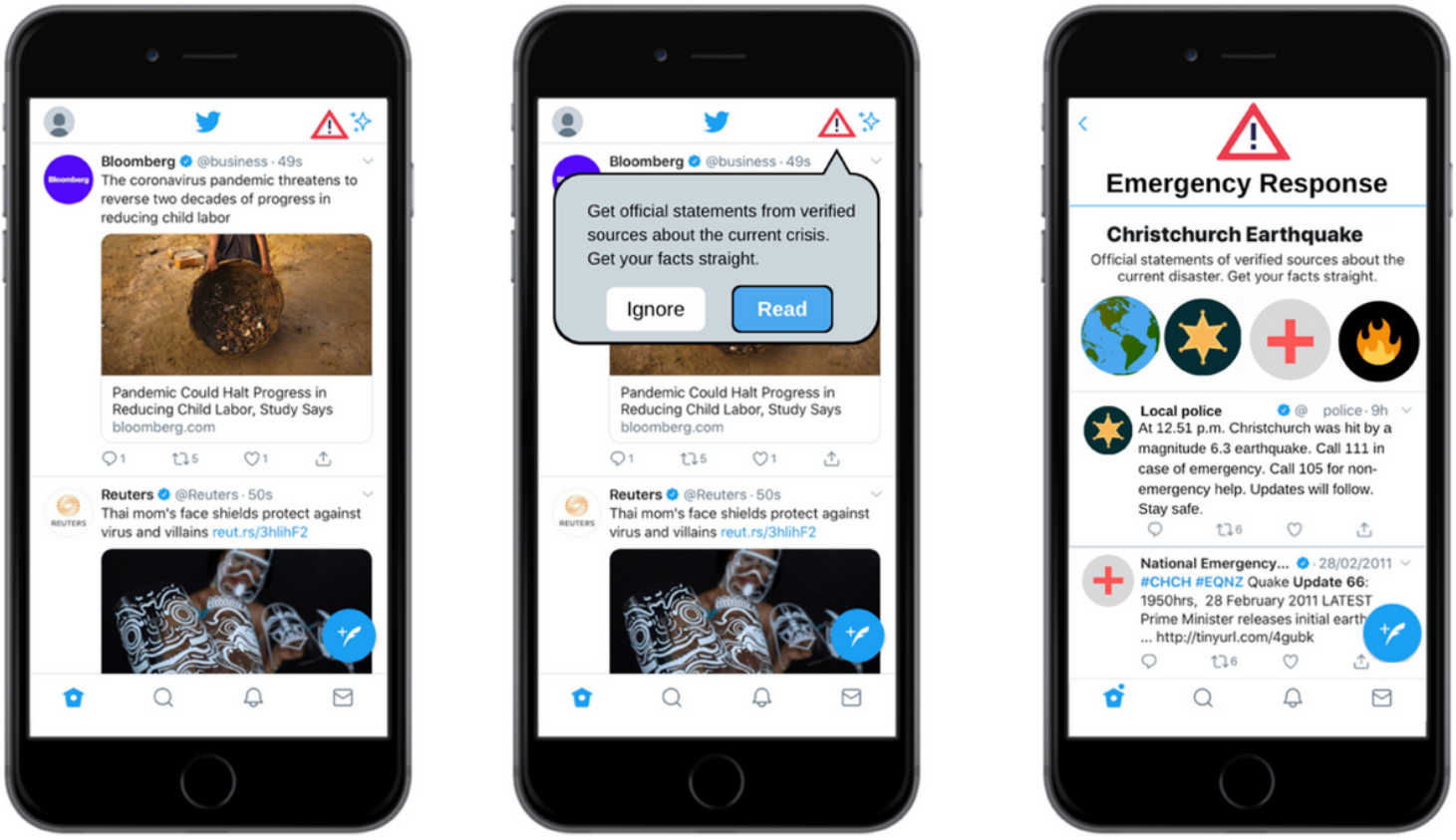

Figure 2

Alternate feed with trustable (local) sources during emergency response

Figure 3

A green verified icon for relevant (local) accounts during disasters