# Designing for affective warnings & cautions to protect against online misinformation threats

Fiona Carroll
Cardiff School of Technologies
Cardiff Met University
Llandaff Campus,
Western Avenue, Cardiff, CF5 2YB
Wales
*fcarroll@cardiffmet.ac.uk*

Bastian Bonkel
Cardiff School of Technologies
Cardiff Met University
Llandaff Campus,
Western Avenue, Cardiff, CF5 2YB
Wales
*bbonkel2@cardiffmet.ac.uk*

**Social media's affordance for misinformation is compromising the glue that holds us and our society together. By influencing and manipulating our human behaviour particularly the decisions we make and opinions we form, it is polarising our existence in not only the virtual but also the physical world in which we live. Yet, despite being aware of the destructive nature of misinformation in general, many of us still don't seem to understand/ see the full danger on an individual basis. Hence, as we have witnessed during Covid 19, many people still continue to share this misinformation widely. The authors of this paper feel that there is an urgent need to support people in being more aware of false information whilst online. In this paper, we share thoughts around some of the mechanisms that people currently use to identify misinformation online. In particular, the focus is on a study that explores participant's experiences of ten different visualisation effects on a Facebook page. The findings highlight that some of these initial visualisation designs are more effective than the others in informing people that something is not quite what it should be. Like in the physical world, we propose the design of a set of affective online visual warnings and cautions that we hope can be further developed to fight online misinformation and counter it's current negative influence on society.** *Misinformation, Warning, Caution, Affective, Visualisation effects, Awareness, Perception.*

## 1. INTRODUCTION

Many countries around the world have spent years trying to build up a socially cohesive society. A society that 'works towards the well-being of all its members, fights exclusion and marginalisation, creates a sense of belonging, promotes trust and offers its members the opportunity of upward mobility' (Oecd 2012, p. 17). However, it seems that the Internet and particularly social media have very quickly started to erode this effort. Moreover, social media's affordance for online misinformation is compromising the glue that holds us and our society together. For example, the spread of misinformation during the coronavirus outbreak was rapid and caused huge uncertainty and tensions amongst people. So much so that the British Computing Society (BCS 2020) in their article '11 ways to fight Coronavirus misinformation' advised that bad spelling is a strong signal of misinformation. However, using grammar and spelling as an indicator of misinformation is becoming less and less useful. As research shows 'digital misinformation thrives on an assortment of cognitive, social, and algorithmic biases and current countermeasures based on journalistic corrections do not seem to scale up' (Ciampaglia 2018, p.147). In reality misinformation that has bad grammar and spelling is likely to increase people's vulnerability to the more sophisticated misinformation attempts. In this paper, we share thoughts around some of the mechanisms that people currently use to identify misinformation online. In particular, the focus is on participant's experiences of ten different visualisation effects on a Facebook page. The aim is to support people in taking more notice of potential misinformation threats. The following sections explore how we might enable people (through visual supports- warnings and cautions) to make the right decisions to counteract the spread of misinformation.

## 2. WHAT IS TRUE AND WHAT IS FALSE?

Our lives today are inextricably tied to the Internet and from this the acquisition of data. While this is empowering many of us, it is also proving to be very harmful especially as it is now more difficult than ever to decipher what is true and what is false in all this data. As Zhou and Zhang (2007, p.1) describe misinformation is the 'transmission of distortions or falsehoods to the audience'. It is distinct from disinformation where false information is spread with the intent to harm, misinformation is the unintentional spread of false

information. Needless to say, both have become such a common part of our digital media environments that it is compromising the ability of our societies to form informed opinions (Fernandez and Alani 2018). Furthermore, it is people's emotions that has become the driving force for much of the widespread of misinformation. As a result it is becoming more and more difficult to centrally control. In detail, content that evokes higharousal positive (awe) or negative (anger or anxiety) emotions is more viral (Berger and Milkman 2012). The authors of this research are interested in the emotional hook of misinformation. In particular, how we can engage the affective through designs to alert (i.e. warn and/ or caution) against misinformation.

## 3. MISINFORMATION, TRUST AND EMOTION

Trust and distrust have been considered as polar opposite constructs (Mal et al. 2018). Trust is the 'willingness to take a risk' and the level of trust is an indication of the amount of risk that one is willing to take (Mayer et al. 1995, p.1). Trusting is the inclination of a person 'A' to believe that other persons 'B' who are involved with a certain action will cooperate for A's benefit and will not take advantage of A if an opportunity to do so arises (Ben-Ner and Halldorsson 2010). In their paper, Schul et al. (2008) see the state of trust as being associated with a feeling of safety, they assume that a state of distrust is the mental system's signal that the environment is not normal, things may not be as they appear. Hence, individuals sense they should be on guard/ careful. If the environment is as it normally is and things really are as they appear to be, then the individuals see no reason to refrain from doing what they routinely do (Schul et al. 2008).

In terms of the affective, Martel et al. (2020) found both correlational and causal evidence that reliance on emotion increases belief in fake news. Furthermore, Greenstein and Franklin (2020, p.1) found the suggestibility for false details increased with anger. In attempt to counter this, the authors of this paper aim to use emotions to alert people to misinformation. As Kaiser et al. (2020) highlights, disinformation warnings can – when designed well – help users identify and avoid disinformation. Moreover, Bhuiyan et al. (2018) developed 'FeedReflect' which is a browser extension that nudges users to pay more attention. It uses reflective questions to engage people in news credibility assessment on Twitter. Other research (Lutzke et al. 2019) highlights the potential of simple interventions to prime critical thinking and slow the spread of fake news on social media platforms. As Fazio (2020, p.1) aptly states, it is about 'adding "friction" (i.e. pausing to think) before sharing can improve the quality of information shared on social media'. Supporting that, Pennycook et al. (2020) present results that show how simple and subtle reminders may be sufficient to improve people's sharing decisions regarding information about COVID-19. Therefore improving the accuracy of the information about COVID-19 on social media.

## 4. STUDY

This study took place at Cardiff Met University in July 2020. Its aim is to give some insight into individuals' perception of misinformation. In particular, to probe participant's experiences of ten different visualisation effects on a Facebook page in order to determine which afforded the most effective alert to the threat of misinformation.

### 4.1. Participants

Five hundred and thirty-two participants from the ages of 18 to 74 years completed the study. These included two hundred and seventy females and two hundred and sixty-two males. The majority of participants were from the age range 35-44 years old (one hundred and twenty-five participants). Also most participants (one hundred and sixty-one females and two hundred and four males) were 'employed for wages'. Others included homemakers, students, retired, self employed, out of work and looking for work, out of work but not looking for work, those unable to work, military and other. All participants (over eighteen years old and internet users) were globally recruited through the Dynata Insights Platform.

### 4.2. Methods & Procedure

The study consisted of four main parts. The first part was to probe participants around the concept of misinformation. To avoid priming, we asked participants if they thought it is easy to identify 'something' online that is not quite right (i.e. not quite as it should be)? The second part of the study was focused on gathering data on participant's thoughts and feelings on an image of an authentic Facebook page rendered ten times with a different visualisation effect (see fig.1). On each image, the visualisation effect was randomly applied to one of the three Facebook posts on the page. These ten effects (see fig.1) were based on designs from earlier studies (Carroll et al. 2018), (Carroll et al. 2020).

These included the different use of colour to **block**, **highlight** and **censor** the text on the Facebook post. They also included different explorations of the visual acuity of the text on the Facebook post: **blur**, **convolve**, **erode**, **fog**, **noise** and **wishy**. Finally, a more literal representation of a threat through broken **glass** over the text on the Facebook post was also investigated. The emphasis of the third part of the study was on which visualisation effect was the most effective in making participants more aware that something is not quite as it should be. Finally the last part of the study was interested in probing participant's opinions of what they think

needs happen with regards to protecting themselves against misinformation threats online. The study took approximately 20-30 minutes in duration. It was conducted using the Qualtrics online survey software and open-ended questionnaire questions were used to collect the data. The Ethics Board of Cardiff Met University approved the study methods and procedure and all participants provided online consent for study completion. The following presents a qualitative analysis of the online survey data. **4.3. Data Analysis & Results**

For the first part of the study and in particular, the question: *In your opinion, do you think it is easy to identify 'something' online that is not quite right (i.e. not quite as it should be)? Please elaborate how you would best identify it*, we have applied six phases of thematic analysis (Braun and Clarke 2006). An initial read of the data generated codes such as 'yes; no; true; can; web; sure; hard; source; details; online; site; scams; questions; sense; research; grammar; easy; good; email; check; poor and new'. Building on these codes, themes such as *gut instinct, spelling and grammar, research, review, appearance, source (URL, website, email, padlock), experience of user, too good to be true, expectations, no/ not sure, yes, random* and *didn't understand question*, started to emerge and then time was taken to gather all data relevant to each potential theme. Finally, after a period of reviewing and refinement was undertaken, the following themes were determined to best demonstrate how participants decipher when something is not right online:

- **Intuition**: gut feeling usually makes me feel when something online isn't genuine
  Participant 36.
- **Appearance**: No, it's not that easy, some scams are very sophisticated. Bad spelling or grammar can sometimes be a giveaway, also asking for info a reputable company wouldn't request. Participant 98.
- **Reviews and Research**: I would look at reviews and research everything from different sites first then match the description up.
  Participant 27.
- **Source and Security**: In my opinion it is relatively easy to identify whether 'something' online is not quite right. There are ways to check the authenticity of certain websites and web pages such as anti-virus tracking software. Web browser address bars indicate whether websites or web pages could be trusted or not by symbols signifying whether they could be trusted such as the padlock. Participant 290.
- **User knowledge**: Yes. I have little trouble spotting these things, but I have been using the Internet for many years and am naturally sceptical. Participant 52.
- **Exceeded expectations**: If an offer seems too good to be true or if the advert does not seem professional. Participant 381.
- **Unrealistic demands**: Asking for personal information when it's literally not needed. Participant 97.

Interested to probe this further, it was clear from the data that the appearance of the content and the digital interface design plays an important role in helping one hundred and seventeen test participants to decipher that something was amiss. This theme of appearance included bad spelling and grammar which featured amongst seventy participants individual comments as a strong indicator of misinformation. Furthermore the parallels between these themes and cyber security awareness is important to highlight (especially, when cyber attacks can include various degrees of misinformation).

Part 2 of the study focused on the appearance of the Facebook posts and in particular, the ten visualisation effects (See fig.1). In detail, we asked participants to describe each visualisation effect/ alternation and its possible effect? It is interesting to see that words like danger, red, attention, warning, highlighted and grabbing are used to describe the **highlight** visualisation effect. Similarly, words like unsafe, warning, red, attention, alarming, danger are also used to describe the **block** visualisation effect. Whilst words like blurred, blurry, fuzzy, suspicious, confusing and ignore are used to described the **blur** visualisation effect and the **glass** effect is being described with words such as broken, cracked, smashed, confusing and annoying.

Moreover when we asked the question about which visualisation effect made them question the validity of what they were reading. The **blur** visualisation effect was more effective for women whilst the **block** visualisation effect for men. When probed about which visualisation effect made them feel uncomfortable, it is clear from the data that the **fog** visualisation (female 51% and male 31%) was the effect that most found uncomfortable. The majority of participant felt that they would ignore the **convolve** effect because it didn't really captivate or interest them. The **block** visualisation effect was of most interest to participants (29% and 22%); it was the visualisation that made them want to know more about why it was altered.

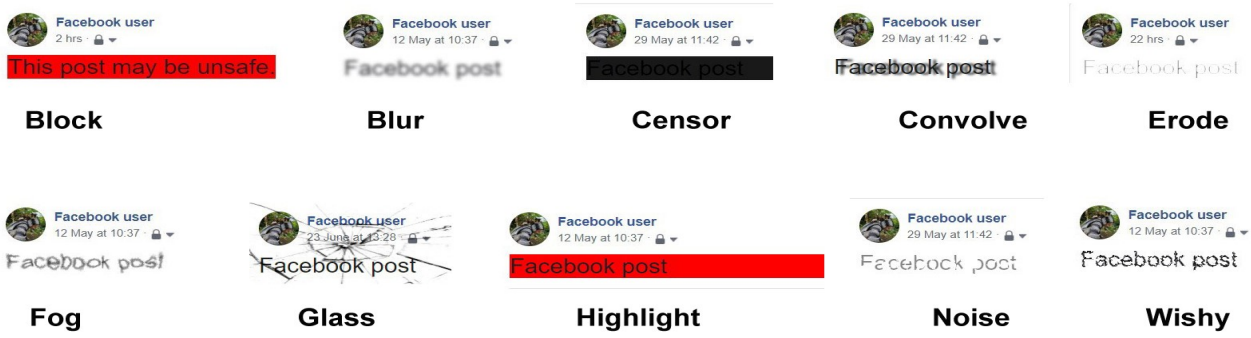


***Figure 1:*** *Ten different visualisation effects*

When asked if they felt nervous or calm looking at these visualisation effects, most participants (one hundred and thirty two participants) felt that the **block** and then **glass** (one hundred and thirty one participants) made them more nervous. When asked which effect made them more alert, the **block** visualisation (one hundred and thirty four participants) showed the highest number of participants that felt it made them more alert (see table 1).

***Table 1:*** *Summary of semantic data captured from Block visualisation effect*

| Semantics (Block) | 1 | 2 | 3 | 4 | 5 | Mean | Mode |
|---|---|---|---|---|---|---|---|
| Nervous (1) to Calm (5) | 35.06% | 23.17% | 25.91% | 9.45% | 6.40% | 2.289634 | 1 |
| Relaxed (1) to Worried (5) | 1.31% | 4.10% | 22.40% | 32.81% | 39.38% | 3.716463 | 4 |
| Attentive (1) to Inattentive (5) | 14.52% | 22.85% | 36.69% | 14.52% | 11.42% | 2.268293 | 1 |
| Unaware (1) to Alert (5) | 0.15% | 1.03% | 14.33% | 35.27% | 49.23% | 4.14939 | 5 |
| Confident (1) to Not Confident (5) | 2.38% | 5.11% | 33.33% | 20.81% | 38.36% | 3.457317 | 3 |

Part 3 of the study was primarily concerned with examining which of the ten visualisations effects was most successful in alerting/ making the participate more aware. In detail, we asked participants to rank each visualisation effect in order of which one makes them most aware that something is not quite as it should be? (1 [top] = Most aware and 10 [bottom] = Least aware). The **block** visualisation effect featured the most ranked at 1.

Finally, for part four of the study, participants were asked what they felt needed to happen online for them (and people in general) to care more about the validity and safety of the online experience. The findings strongly show that people need to be made more aware of what is happening. The frequency of words such as warnings (32 times), alerts (21 times), checks (9 times), messages (8 times) highlight that participants feel they need the support to become more aware.

## 5. CONCLUSION

In conclusion, we feel that there is currently a lack of support for people to identify a misinformation threat in the online environment. In the physical world we are provided with a range of techniques to enable us to determine whether something needs to be fully avoided or simply to take heed with. In the online environment, we don't have a set of standards or laws detailing what symbols /signs/ effects that determine what is dangerous or what might afford or connote careful and attentive behaviour.

Moreover, we feel that knowing the difference between the online warning and caution is essential for further online interactions. As an end user, we need to be able to perceive and understand that a caution online indicates a minor risk to ones person if proper safety practices aren't observed. Whilst also, to understand that a warning online is an alert to significant dangers. As this study has started to show, certain visualisation effects can trigger certain feelings around online information. Also, in parallel, people seem to be naturally examining the presentation of their online environments as a means to detect if something is not quite as it should be. This research aims to support this behaviour further by providing end users with a more effective means to identify when something is lacking in integrity online. This particular study is the first in a series of studies to explore the development of effective online warnings and cautions. Similar to the physical world, the aim is to provide people with a system of warnings and cautions to protect them against online threats (including misinformation).

**ACKNOWLEDGEMENTS**

This research was funded by the Welsh Crucible, a consortium of Welsh higher education institutions and the Higher Education Funding Council for Wales (HEFCW). We are very grateful to Dr James Kolasinski, Cubric, Cardiff University who was a collaborator on this research project.

**REFERENCES**

BCS (2020). 11 ways to fight Coronavirus misinformation — BCS.

Ben-Ner, A. and F. Halldorsson (2010). Trusting and trustworthiness: What are they, how to measure them, and what affects them. *Journal of Economic Psychology*.

Berger, J. and K. L. Milkman (2012). What makes online content viral? *Journal of Marketing Research*.

Bhuiyan, M. M., K. Vick, T. Mitra, K. Zhang, and M. A. Horning (2018). FeedReflect: A tool for nudging users to assess news credibility on twitter. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*.

Braun, V. and V. Clarke (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*.

Carroll, F., M. Webb, and S. Cropper (2018). Losing our senses online: Investigating how aesthetics might be used to ground people in cyberspace. *IEEE Technology and Society Magazine*.

Carroll, F., M. Webb, and S. Cropper (2020, sep). Investigating aesthetics to afford more 'felt' knowledge and 'meaningful' navigation interface designs. In *2020 24th International Conference Information Visualisation (IV)*, pp. 214–219. IEEE.

Ciampaglia, G. L. (2018). Fighting fake news: a role for computational social science in the fight against digital misinformation. *Journal of Computational Social Science*.

Fazio, L. (2020). Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review*.

Fernandez, M. and H. Alani (2018). Online Misinformation: Challenges and Future Directions. In *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*.

Greenstein, M. and N. Franklin (2020). Anger Increases Susceptibility to Misinformation. *Experimental Psychology*.

Kaiser, B., J. Wei, J. N. Matias, E. Lucherini, J. Mayer, and K. Lee (2020). Adapting Security Warnings to Counter Online Disinformation.

Lutzke, L., C. Drummond, P. Slovic, and J. Arvai´ (2019). Priming critical thinking: Simple interventions limit the influence of fake news about climate change on Facebook. *Global Environmental Change*.

Mal, C. I., G. Davies, and A. Diers-Lawson (2018). Through the looking glass: The factors that influence consumer trust and distrust in brands. *Psychology and Marketing*.

Martel, C., G. Pennycook, and D. G. Rand (2020). Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications*.

Mayer, R. C., J. H. Davis, and F. D. Schoorman (1995). AN INTEGRATIVE MODEL OF ORGANIZATIONAL TRUST. *Academy of Management Review*.

Oecd (2012). Perspectives on Global Development 2012 Social Cohesion in a Shifting World Executive summary. *Development*.

Pennycook, G., J. McPhetres, Y. Zhang, J. G. Lu, and D. G. Rand (2020). Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*.

Schul, Y., R. Mayo, and E. Burnstein (2008). The value of distrust. *Journal of Experimental Social Psychology*.

Zhou, L. and D. Zhang (2007). An ontologysupported misinformation model: Toward a digital misinformation library. *IEEE Transactions on Systems, Man, and Cybernetics Part A Systems and Humans*.