Toward Designing Effective Warning Labels for Health Misinformation on Social Media

Huma Varzgani
Worcester Polytechnic Institute
hvarzgani@wpi.edu

Nima Kordzadeh
Worcester Polytechnic Institute
nkordzadeh@wpi.edu

Kyumin Lee
Worcester Polytechnic Institute
kmlee@wpi.edu

Abstract

Health misinformation on social media has become a major threat to users. To alleviate this issue, platforms such as Twitter have started labeling posts considered as misinformation to warn users. However, the effectiveness of such labels on user perceptions and actions are not clear, as it has not yet been examined by researchers in prior studies. We aim to address this gap through a model, which draws upon concepts from color theory and construal level theory and focuses on the impact of three misinformation label characteristics: background color of the label, abstractness of the message, and assertiveness of the message language. We propose that the effectiveness of these warning labels will lead users to verify, avoid using, and avoid sharing such labeled posts on social media. This paper provides important theoretical contributions and aids policymakers and platform providers by offering insights on what motivates users to take protective actions.

Keywords: Warning labels, misinformation, color, abstractness, assertiveness.

1. Introduction

Over the past decade, social media platforms have been widely used for seeking and providing health-related information. Results of a study conducted by Alhusseini, Banta, Oh, and Montgomery (2020) using data from National Cancer Institute's Health Information Trends Survey (HINTS) showed that 67.5% of US adults used electronic means to find health information online (Alhusseini et al., 2020). As the Internet is easily accessible to many, they are prompted to seek health information online. It can also be an appealing source due to the lack of access to affordable healthcare in some countries (Bhandari et al., 2014). Nonetheless, while using social media for sharing and seeking health-related information has benefits, it may lead to generation and propagation of misleading information, fake news, and pseudoscientific advice, which are collectively known as misinformation. From politics to health, social media has become a platform where misinformation can take root and spread; however, the consequences of misinformation in the healthcare context can be more serious as it could "generate feelings of apathy, confusion, and mistrust" due to which individuals might make decisions such as avoidance of vaccinations and forgoing recommended therapies or treatments (Sylvia Chou & Gaysynsky, 2020) that can be detrimental to their health (Sylvia Chou et al., 2020). Social media users may share health misinformation either intentionally to generate conspiracy theories (Halpern et al., 2019) or unintentionally as they trust the source that initially shared it. Studies show that 23.8% of hundred and one videos posted on YouTube related to zika virus (ZIKV) contain misleading information (Bora et al., 2018) and 23% of the videos related to COVID-19 are misleading, 30.4% of which deemed 'false', 52.2% contain unproven theories, and 17% contain conspiracy theories (Andika et al., 2021). Sharing fake or misleading information may entail negative consequences for individuals and society. Hence, it is important to understand how health misinformation propagation via social media can be mitigated.

To minimize the spread of misinformation, platforms such as Twitter have started placing warning messages with corrective information as labels attached to the posts that are deemed to be misleading. Though similar to information security warnings, where users are alerted of any security threat discovered that may damage their computer systems or breach their personal information (Zaaba et al., 2021), misinformation warning labels specifically focus on mitigating the spread and use of misinformation that directly impact the health of users. For example, during the COVID-19 pandemic, Twitter used labels to warn users about false or misleading tweets related to topics such as vaccination (Sharevski et al., 2021). In this way, Twitter intended to keep

URI: https://hdl.handle.net/10125/102646
978-0-9981331-6-4
(CC BY-NC-ND 4.0)

people from spreading or consuming falsehood. Nonetheless, the efficacy of those warning labels has been questioned by researchers and practitioners, as it has drawn criticism for not being effective from keeping people from spreading misinformation. One of the reasons, for example, was the color of the labels, which was blue that blended with Twitter's color scheme and could not catch the users' attention (Ortutay, 2021). Accordingly, it is not clear how those labels should be designed effectively and what label characteristics can impact users' perceptions and actions regarding labeled misinformation. Twitter has been experimenting with different types and forms of labels; however, more empirical studies are needed to unravel the role of different label characteristics in enabling users to verify misinformation and nudging them to not use or share misinformation. In this paper, we theorize the impact of three misinformation label characteristics on users' willingness to engage in protective actions including (1) verifying, (2) avoid using, and (3) avoid sharing labeled social media posts that contain misinformation. The three label characteristics include background color, concreteness of the warning message (i.e., the degree to which a message is specifically and tangibly described), and assertiveness of the warning message (i.e., the extent to which the language used in a message is clear, direct, and non-suggestive). These three label characteristics serve as our focal influencing factors as according to previous research, different characteristics of warning messages including content of the message, format in which the information is presented (Stewart & Martin, 1994), and color of the warning messages (Silic & Cyr, 2016) can have different cognitive effects on the users.

Color has been used in various contexts to draw attention or to stimulate certain emotions among people (Kaya & Epps, 2004). Different colors trigger different psychological responses, due to which colors are used to arouse the desired emotions (Kaya & Epps, 2004). In particular, the temperature of colors affects people's decision-making and behavior. Warm colors, such as red, orange, and yellow, are considered more arousing and invigorating than cool colors, such as blue, purple, and green (Torres et al., 2020). Based on this reasoning and drawing on the color theory, we, in this study, theorize that warm background colors in warning labels are more effective and persuasive for users to engage in protective actions against misinformation.

Furthermore, misinformation warning messages can be considered as fear appeals, which can be defined as messages that may conjure up "a potential emotional response to threats" (Cauberghe et al., 2009), where threat is an external stimulus (Witte, 1992). When individuals face a threat that is depicted as, and believed to be, significant and relevant, it evokes fear, which is an unpleasant feeling (Zhang & Zhou, 2020). This feeling encourages them to take protective actions against that threat. We propose that with the use of appropriate fear language in the context of warning messages, social media platforms can drive users to verify misinformation and to avoid using and sharing posts that are labeled as misinformation. The two aspects of fear language that we theorize in this study are concreteness and assertiveness of warning messages.

While several studies have been conducted to determine the effectiveness of fear appeals in public health campaigns (Laroche et al., 2015; Lewis et al., 2007; Meadows, 2020; Zhang & Zhou, 2020), little research has been carried out to understand how fear appeals in the form of warning messages can impact users' intentions to verify, use, and share health-related (mis)information on social media. Previous literature was more focused on the effects of fear appeals on variables such as self-efficacy and response efficacy (Cauberghe et al., 2009; Lennon & Rentfro, 2017; Ruiter et al., 2014; Tannenbaum et al., 2015; Witte & Allen, 2000), whereas it is important to understand how, in the context of health misinformation on social media, fear appeals can be used to help users identify misinformation and take protective actions against them. In this study, we develop a theoretical model to enhance understanding of this phenomenon. More specifically, we draw upon, integrate, and extend the color theory and construal level theory to theorize the role of warning label color as well as concreteness and assertiveness of warning messages in this context. Based on color theory, the use of appropriate color can evoke certain feelings, such as fear, within individuals. Similarly, based on construal level theory, message content and language (concreteness and assertiveness) can also be used to evoke the same feelings of fear, which can encourage people to take protective actions. While these theories have been used in contexts such as information security to understand users' protective behaviors against security threats (Mady, 2017; Schuetz et al., 2020), to the best of our knowledge, no study has used these theories to explain social media users' responses to misinformation warning labels, particularly in the public health information domain. Moreover, prior studies have not tried to examine a combination of different characteristics of misinformation warning messages, particularly color and fear appeal language, to understand which factors are more critical in designing warning messages. For example, previous research has proven that some signal words are associated with colors (Braun et al., 1995; Braun &

Silver, 1995; David Leonard, 1999; Kline et al., 1993), but research about which colors can be used to effectively warn users about misinformation is still lacking. This study takes an important theoretical step to fill these gaps in the literature and paves the way for future empirical studies that can examine this phenomenon. Thus, we seek to address the following research questions:

RQ1: How does the color of a misinformation warning label impact the verification, using, and sharing of the labeled content on social media?

RQ2: How do the concreteness and assertiveness of the fear appeal language in a misinformation warning message impact the verification, using, and sharing of the labeled content on social media?

2. Literature Review

2.1. Color Theory

Color has much more importance and relevance in our everyday life than it is deemed. Its use is not limited to aesthetics as it can largely affect users' decision making capabilities in different contexts as different colors can carry certain meanings (Elliot & Maier, 2007). The color pink, for example, is usually associated with females and the color blue with males (Chiu et al., 2006). Colors impact our decision making capabilities by affecting our cognitive system, perception, psychological and emotional reaction, as well as our behavioral intentions (Elliot & Maier, 2007; Kaya & Epps, 2004). The psychological influence that colors have on us usually takes place without any conscious intention (Elliot & Maier, 2007). Without us being consciously aware of it, our decisions are sometimes guided by the color that we interact with (Silic et al., 2016); for instance, drivers stop at the signal when it turns red as the color red indicates that there is danger ahead (Silic & Cyr, 2016).

There is more to the meaning of color, according to the color theory; the temperature of the color plays an essential role in impacting our cognitive thinking capabilities due to the varying arousal effects that different colors have on us (Kurt & Osueke, 2014). The color wheel is split into two broad groups of colors, warm and cool colors. Colors closer to the red end of a visible spectrum are referred to as warm, which includes colors like orange and yellow, and colors closer to the blue end are considered cool like purple and green (Bailey et al., 2006). Warm colors also have longer wavelengths than cool colors. Warm and cool colors have different levels of appeals (Silic et al., 2016). According to Elliot and Maier (2007), warm colors create more arousal than cool colors as cool colors elicit a calming and relaxing effect. Warm colors are considered dominant and more stimulating (Kaya & Epps, 2004). As indicated by Singh (2006), colors like red and yellow are typically used in marketing to gain customer attention, as these colors can stimulate certain feelings, which in this context is appetite.

Due to the effects that colors have on our emotions and decision making, they are used in various ways to draw attention (Dzulkifli & Mustafar, 2013). Specifically, color is one of the qualities that should be considered when designing warnings. Several studies have been conducted on the importance of colors in warning messages. Results from a study by Kline et al. (1993) exhibited that warning labels with color were perceived as more "readable" and "hazardous" than labels that were black and white (Kline et al., 1993). Braun and Silver (1995) suggest that color, along with a written warning, has a higher effect on perception of hazards and warning conformity (Braun & Silver, 1995). Certain colors are related with certain levels of risk associated with a hazard (David Leonard, 1999). In their study, Braun et al. (1995) discovered that signal words in red and orange produced higher ratings of perceived hazard than signal words in blue, when a group of undergraduates were asked to rate the perceived hazard of several signal word and color combinations printed in specific hazard colors (Braun et al., 1995). Despite the fact that several studies have looked into the impact of color on peoples' judgements and decision making, to our best knowledge, no study has specifically theorized the role of color in misinformation warning messages on social media; this study theoretically addresses this gap in the literature.

2.2. Fear Appeal

The use of fear appeals has been found to be effective in inducing fear among individuals, especially when used in the contexts of health warning and public service announcements (PSAs) (Hastings et al., 2004). As Meadows (2020) investigated the effect of high threat and neutral threat health warnings in the context of promoting skin cancer prevention behaviors among students, he found that fear has considerable effect on individuals' assessment of risk (Meadows, 2020). According to a meta-analysis conducted by Witte and Allen (2000) on fear appeals, there are three key independent variables associated with fear appeals: fear, perceived threat, and perceived efficacy (Witte & Allen, 2000). The results of these variables fall under two categories. First, how individuals are

open to accept the recommendation provided in the message which includes attitudes, intentions, and behaviors. Second, how the outcomes are related to rejection of the message, which includes defensive avoidance, reactance, and denial (Witte & Allen, 2000). In another meta-analysis, Tannenbaum, Helper et al. (2015) discussed that when fear appeal messages are accompanied with efficacy, individuals are more likely to act on the recommended action (self-efficacy) or perform an action that will result in desirable outcomes (response-efficacy), as efficacy in a message assures the recipients that they are capable of performing the recommended actions (Tannenbaum et al., 2015).

The language used to promote fear appeals has also immense effect in persuading individuals to take preventive actions. Previous studies have shown that the use of fear appeals is efficacious when they are used moderately – that is, messages with high intensity of fear (i.e., using very strong and sensationalized language in depicting the severity of risk) might not give impactful results. Albarracín, Gillette et al. (2005) examined the long-term efficacy effect of high intensity fear-inducing arguments about HIV counseling and testing as well as encouraging people to gain knowledge about HIV transmission and prevention. The results of the study revealed that the use of high intensity fear inducing arguments increased the risk perception but proved to be ineffective in promoting HIV-relevant learning (Albarracín et al., 2005). In separate studies, Rhodes (2015), Cauberghe et al. (2009), and Lewis et al. (2007) found that when college students viewed PSAs about safe driving with moderate level of fear, they were effectively persuaded to drive more slowly and carefully (Cauberghe et al., 2009; Lewis et al., 2007; Rhodes, 2015). Thus, it is critical to consider the intensity of emotional aspects of the message when crafting a warning message (Rhodes, 2015; Witte & Allen, 2000).

As summarized by Ruiter, Kessels et al. (2014), strengthening self-efficacy, promotion of response efficacy, awareness of susceptibility, and suggesting that the threat is severe in an emotional way are vital elements of fear appeal messages to motivate risk reduction behaviors (Ruiter et al., 2014). In a similar manner, other language aspects such as message concreteness and message assertiveness are also important factors, which can impact an individual's action. Construal level theory framework can be used to explain how message concreteness can strengthen fear appeals and how message assertiveness can encourage individuals to take protective action.

2.3. Construal Level Theory

Construal level theory explains how individuals perceive psychologically distant objects or events (Trope & Liberman, 2010). The mental representation of how these distant objects or events are perceived can be defined as construal (Wiesenfeld et al., 2017) and the psychological distance of these construals can be described as when, where, and to whom the perceived event can occur (Trope & Liberman, 2010). For example, someone is working on an important project on their computer and it suddenly shuts down. They would begin interpreting the situation and find a possible explanation for the situation, why the computer shut down, whether the computer's battery ran out, was it due to a virus attack, and how the computer could be turned on again. This interpretation of the situation is called construal. According to cognitive psychology, people use cognitive structures such as schemas and categories to make sense of information or situations (Rosch, 1975) and the cognition attached with these structures can be described as abstract or concrete. An abstract cognitive structure is less comprehensive and focused, whereas a concrete cognitive structure is more focused (Wiesenfeld et al., 2017). Correspondingly, construals can also be described as abstract or concrete. Abstract construals are relatively more broad and general compared to concrete construals, which are more specific and detailed (Wiesenfeld et al., 2017).

Similar to how individuals form a mental representation of abstract and concrete construals, they can make a mental representation of abstract and concrete fear appeals (Schuetz et al., 2020). Schuetz et al. (2020) applied this concept in an information security setting and explained how abstract and concrete fear appeals differ in terms of characteristics and effects. Accordingly, just like abstract construals, abstract fear appeal messages are more generic (e.g., "phishing is dangerous"), and similar to concrete construals, concrete fear appeals are more specific (e.g., "phishing asks for your credit card information") (Schuetz et al., 2020). Across three experiments, they found that concrete fear appeal messages were more effective in influencing intentions and behaviors than abstract fear appeal messages by improving fear efficacy and stimulating threat severity, threat vulnerability, fear, and protection motivation behavior.

Construal level theory also suggests that individuals are motivated to make better decisions when the distance matches the nature of the problem (Liberman et al., 2007). Concrete construals have smaller temporal distance, which can be described as how far an imagined future event is from present (Bar-

Anan et al., 2006). As concrete fear appeals explicitly describe the problem, they imply that action must be taken in the near future, thereby, decreasing the temporal distance (Nan, 2007) and successfully influencing individuals to adopt coping strategies. In this vein, a study that assessed the effects of temporal framing in messages about smoking-related diseases and recommended quitting smoking found that the near-future frame (i.e. concrete) resulted in greater perceived susceptibility to the risk depicted in the message and greater intention to quit smoking (Kim & Kim, 2018).

Following the construal level theory , the "how" and "why" aspects of fear appeal message can also be explained as *argument nature*. Schuetz et al. (2021) proposed and validated argument nature as a significant variable to be considered when designing fear appeal messages, as it was found in that study that users were more likely to show the behavioral response when they were suggested how the preventive action could be executed and why it was imperative (Schuetz et al., 2021). When the "how" and "why" aspects of the fear appeal message are described in a more assertive tone or language, it may further enhance the promotion of protective behavior. Assertive communication is viewed as an essential component when it comes to patient safety (Omura et al., 2017). Compared to suggestive tone, assertive tone is considered to be more effective in conveying the desired message as individuals are not provided with the option to refuse (Kronrod et al., 2012). However, there are instances where tone and language used in messages might have different effectiveness in different settings, but in more pro-social settings, assertive language has proven to be more useful (Grinstein & Kronrod, 2016). In line with the construal level theory, assertive language implies that psychological distance between now and future (where the threat lies) can be smaller, and thus demands action to be taken. For example, Schuetz and Lowry et al. (2021) found that in an information security context when temporal distance was proximate, threat vulnerability and self-efficacy were higher, enforcing protective actions (Schuetz et al., 2021).

By inhabiting the concepts of construal level theory to explain the impact of message abstractness and message assertiveness, along with the color theory, we develop our theoretical model and present three hypotheses in the following section.

3. Theoretical Model

In this section, we theorize why the warning labels that are presented in warm color backgrounds (H1), include concrete messages (H2), and use assertive language (H3) are expected to be more effective in driving users to take protective actions against the misinformation threat on social media. Our theoretical model is demonstrated in Figure 1.

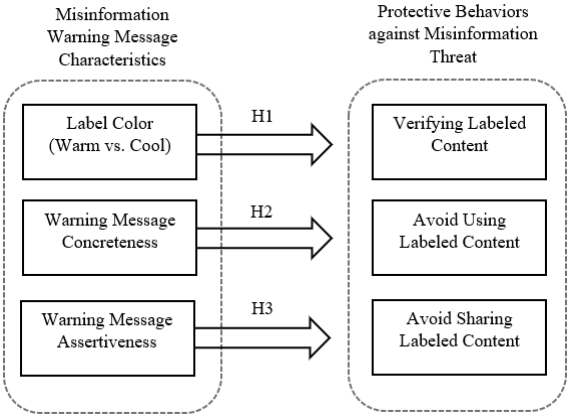

Figure 1. Theoretical model.

As discussed earlier, colors play a vital role in our lives. They not only brighten our lives but also play a huge part in swaying our thoughts and invoking reactions and emotions, which can influence our perspective and behaviors. Prior research shows that warm colors are considered more arousing and stimulating than cool colors (Elliot & Maier, 2007; Kaya & Epps, 2004; Singh, 2006). Colors with higher temperatures are considered to be more vigorous and thus, more impactful in exhilarating certain emotions, which in turn affect our cognitive thinking abilities (Kaya & Epps, 2004; Silic & Cyr, 2016; Silic et al., 2016). Moreover, some colors have a certain level of risk associated with them. For example, red is typically associated with danger, orange with warning, and yellow with caution (David Leonard, 1999). As prior research suggests, warm colors are more commonly used to exhibit warnings. Compared to cool colors, warm colors catch user attention more easily (Singh, 2006), prompting users to pay attention to the warning label and read the message. Accordingly, we argue that the use of warm colors will be more effective in evoking protective behavior against the misinformation threat among users. Furthermore, as warm colors are associated with warnings (David Leonard, 1999), they can effectively help in warning social media users and promoting protective behavior. This can encourage users to verify the social media content (i.e., the tweet) with a warning label and

accordingly, not follow or share the content. Following this explanation, we hypothesize :

H1: Warm colors in warning labels are more effective than cool colors in driving users to engage in protective actions against misinformation on social media.

H1a: Warm colors in warning labels are more effective than cool colors in driving users to verify labeled social media contents.

H1b: Warm colors in warning labels are more effective than cool colors messages in driving users to avoid using labeled social media contents.

H1c: Warm colors in warning labels are more effective than cool colors in driving users to avoid sharing labeled social media contents.

Along with color playing an important role in eliciting feelings of fear when users first interact with a message, its content can also influence users' perceptions. According to the construal level theory, fear appeal messages can have an influence on forming construals in users' minds, and thus impact their decision making and subsequent behaviors (Schuetz et al., 2020). Research shows that when events are explained more concretely, it is easier for users to focus on the incidental features, which may alter their feelings and thoughts and may motivate their risk reduction behavior due to the persuasive nature of concrete messages (Wiesenfeld et al., 2017). Compared to abstract fear appeal messages, concrete fear appeals describe distant events more truthfully, and hence in more negative terms, which increases threat severity and strengthens intentions to follow the provided recommendation (Klohn & Rogers, 1991). However, the intensity of negativity or fear evoked must be moderately used to keep it impactful (Cauberghe et al., 2009; Lewis et al., 2007; Rhodes, 2015). In line with this, Klohn and Rogers (1991) examined different aspects of the severity of a health threat to a sample of female subjects who were at risk for osteoporosis. The subjects did not consume the adequate amount of calcium daily neither did they participate in daily exercises. The study resulted in motivating them to take protective actions when they were displayed with messages with high severity (Klohn & Rogers, 1991). Accordingly, we argue that, in the context of misinformation on social media, concrete warning messages, which are more detailed and specific than abstract messages (Schuetz et al., 2020), can more successfully invigorate protective behaviors. This is because concrete messages inform users more effectively about the issue at hand and also provide them with what steps to take to prevent the threat of misinformation. This likely amplifies user intention to verify labeled social media content and also discourage them from using or sharing the labeled content. Because concrete messages, just like concrete construals, illicit that a threat is proximate (Trope & Liberman, 2010), users will be motivated to take action to avoid that threat. Thus, we hypothesize:

H2: Concrete warning messages are more effective than abstract warning messages in driving users to engage in protective actions against misinformation on social media.

H2a: Concrete warning messages are more effective than abstract warning messages in driving users to verify labeled social media contents.

H2b: Concrete warning messages are more effective than abstract warning messages in driving users to avoid using labeled social media contents.

H2c: Concrete warning messages are more effective than abstract warning messages in driving users to avoid sharing labeled social media contents.

The language used to present the content of the message can also affect the users' perceptions of fear and compliance. Compared to suggestive language, which is considered more indicative, assertive language, which is more commanding in tone, induces quicker compliance (Kronrod et al., 2012). Although assertive communication style is considered more imposing, it still conveys the message in less aggressive and humiliating manner (Pipaş & Jaradat, 2010).

Furthermore, messages that provide recommendations in assertive tone demand immediate action, and thus demonstrate higher compliance (Kronrod et al., 2012). Looking into this from a marketing perspective, brands sometimes use assertive and forceful language to promote their products to fasten their consumers readiness to comply; Nike's slogan "Just do it" is an example of this notion (Wang & Zhang, 2020). Kronrod, Grinstein, and Wathieu (2012) studied the language used in environmental slogans and found that 57% of those slogans were assertive in nature (Kronrod et al., 2012). Similar to these slogans, fear appeal messages written in an assertive language demand immediate action by showing little distance (temporal) between present and the future, where the future represents the consequences of not taking that action. Users are more distant from performing the protective action if they are not forced to do so; accordingly, we posit that temporal distance is smaller when assertive language is used in fear appeal messages instead of suggestive language.

In addition, when the "how" and "why" aspects of the argument nature is explained more concretely and assertively, users are forced to consider "How am I at risk?", "How can I protect myself?", "Why do I need to avoid this threat?", and "Why should I perform this response?" (Schuetz et al., 2020). Accordingly, by

explicitly narrating how to take preventive action, assertive communication styles can induce self-efficacy and stimulate users to follow the preventive recommendation. Aligned with the prior literature, we argue that users will be more inclined toward verifying the information before using or sharing it. Thereby, we propose:

H3: Assertive language in warning messages is more effective than suggestive language in driving users to engage in protective actions against misinformation on social media.

H3a: Assertive language in warning messages is more effective than suggestive language in driving users to verify labeled social media contents.

H3b: Assertive language in warning messages is more effective than suggestive language in driving users to avoid using labeled social media contents.

H3c: Assertive language in warning messages is more effective than suggestive language in driving users to avoid sharing labeled social media contents.

4. Discussion

Due to the proliferation of health-related data, self-diagnosis among online users has become very common. People use the resources available on the internet to investigate their health conditions. On average, 15.49% look up the internet about symptoms that could be related with their medical condition before receiving a medical diagnosis from a medical provider (Hochberg et al., 2020). This has become a global challenge also known as "infodemic" – the spread of too much information, both accurate and inaccurate, in the physical and digital environment (Eysenbach, 2020) . For example, a study conducted at Fondazione Bruno Kessler Institute examined 112 million social media posts related to COVID-19 and found that more than 40% of the social media posts came from unreliable sources and 40% of those posts were circulated by bots (Lupi, 2020). This is major threat to social media users, which should be addressed by policymakers and platform providers. In this paper, we theoretically discussed the important characteristics of the warning message labels that are used by social media platforms to warn users about the health-related posts that are regarded as misleading or inaccurate (Sharevski et al., 2021). As the effectiveness of such labels have not been examined theoretically in prior studies, we addressed this gap by theorizing the impact of three misinformation label characteristics on users' willingness to employ protective actions. The three label characteristics include label's background color, abstractness of the message, and assertiveness of the message language. Drawing on the color theory and construal level theory, we developed a model to theorize the impact of these three characteristics on the effectiveness of warning labels in leading users to verify, avoid using, and avoid sharing labeled posts on social media.

As the current literature on color theory suggests, colors can be utilized to make users respond in a certain way, as colors can trigger different psychological responses (Kaya & Epps, 2004). Not only are colors useful in catching attention, but also, they aid users in decision making. This paper used and extended color theory in the context of labeling health misinformation on social media. Moreover, we built on the notion of fear appeals combined with the construal level theory to theorize the effectiveness of the two aspects of fear appeal language including concreteness and assertiveness of warning messages. We argued that fear appeal messages with low level construals are more specific and concrete and they clearly define the temporal and psychological distance of the construal (Wiesenfeld et al., 2017). We proposed that the more concretely the message is displayed, the more likely it is that users will consider the recommended protective actions (Schuetz et al., 2020; Schuetz et al., 2016; Witte & Allen, 2000). In this way, we theoretically illuminated the role of fear appeal messages in terms of high-level and low-level construals in promoting protective behaviors in the realm of health misinformation on social media.

Although in this paper we did not test the proposed theoretical model, it can serve as a steppingstone for an empirical study (or a series of studies), as researchers in prior research have followed the same theoretical-empirical approach (Kordzadeh & Warren, 2014, 2017). As such, the research team is currently in the process of testing the model through a randomized, between-subject experiment with a 2 (warm color vs. cool color) x 2 (concrete vs. abstract message) x 2 (assertive vs. suggestive language) factorial design to provide evidence on the significance of each factor introduced in the paper and how they can be used to construct more persuasive warning label messages.

Researchers in future studies can also build on, and extend, our theoretical model to further enhance understanding of health-related misinformation and the role of warning labels to mitigate the misinformation threat on social media. Our model currently addresses three important characteristics; however, future studies can examine additional factors such as font, sentiment, and length of warning messages to better understand how they should be designed effectively. Researchers can also employ the same concept in different social media platforms such as Facebook, Instagram, and TikTok to unravel the

platform-specific user behaviors that emerge in responses to warning labels with specific characteristics. In addition, future research can theorize and assess the interactions between the three label characteristics that we discussed in this paper, which can provide more insights into the effectiveness of different forms of warning labels.

5. Conclusion

The main objective of this study was to provide theoretical insights into the critical characteristics of warning label messages that can be helpful in informing users about misleading content and decrease the virality of health-related misinformation. To do so, a theoretical model including three hypotheses, each with three sub-hypotheses, was developed. The background color of the warning label, the concreteness of the message, and the assertiveness of the message were proposed to impact the extent to which users verify, use, and share labeled misinformation. Accordingly, users are more likely to feel obliged to verify the information before using or sharing it if the label attached to it is in warm color and if the message is concrete and assertive. We used concepts from color theory, fear appeal, and construal level theory as the backbone for our model. We also laid out a path for future researchers to better understand how to design warning labels for health-related misinformation on social media effectively.

Acknowledgements

This research was supported by Worcester Polytechnic Institute (WPI), the WPI Business School (grant: Transformative Research and Innovation, Accelerating Discovery - TRIAD), and NSF grant CNS-1755536.

References

Albarracín, D., Gillette, J. C., Earl, A. N., Glasman, L. R., Durantini, M. R., & Ho, M.-H. (2005). A Test of Major Assumptions About Behavior Change: A Comprehensive Look at the Effects of Passive and Active HIV-Prevention Interventions Since the Beginning of the Epidemic. *Psychological Bulletin, 131*(6), 856-897. https://doi.org/10.1037/0033-2909.131.6.856

Alhusseini, N., Banta, J. E., Oh, J., & Montgomery, S. (2020). Understanding the Use of Electronic Means to Seek Personal Health Information Among Adults in the United States. *Cureus*. https://doi.org/10.7759/cureus.11190

Andika, R., Kao, C. T., Williams, C., Lee, Y. J., Al-Battah, H., & Alweis, R. (2021). YouTube as a source of information on the COVID-19 pandemic. *Journal of Community Hospital Internal Medicine Perspectives, 11*(1), 39-41. https://doi.org/10.1080/20009666.2020.1837412

Bailey, R. J., Grimm, C. M., & Davoli, C. (2006). *The Real Effect of Warm-Cool Colors* (WUCSE-2006-17). (All Computer Science and Engineering Research, Issue. http://dx.doi.org/10.7936/K7736P3B

Bar-Anan, Y., Liberman, N., & Trope, Y. (2006). The association between psychological distance and construal level: Evidence from an implicit association test. *Journal of Experimental Psychology: General, 135*(4), 609-622. https://doi.org/10.1037/0096-3445.135.4.609

Bhandari, N., Shi, Y., & Jung, K. (2014). Seeking health information online: does limited healthcare access matter? *Journal of the American Medical Informatics Association, 21*(6), 1113-1117. https://doi.org/10.1136/amiajnl-2013-002350

Bora, K., Das, D., Barman, B., & Borah, P. (2018). Are internet videos useful sources of information during global public health emergencies? A case study of YouTube videos during the 2015–16 Zika virus pandemic. *Pathogens and Global Health, 112*(6), 320-328. https://doi.org/10.1080/20477724.2018.1507784

Braun, C. C., Kline, P. B., & Silver, N. C. (1995, 1-1-1995). The influence of color on warning label perceptions.pdf [Article]. *International Journal of Industrial Ergonomics, 15*(3), 179-187. https://doi.org/http://dx.doi.org/10.1016/0169-8141(94)00036-3

Braun, C. C., & Silver, N. C. (1995). The Interaction of Signal Word and Color on Warning Labels.pdf. *Ergonomics, 38*(11), 2207-2220. https://doi.org/https://doi.org/10.1080/00140139508925263

Cauberghe, V., De Pelsmacker, P., Janssens, W., & Dens, N. (2009, Mar). Fear, threat and efficacy in threat appeals: message involvement as a key mediator to message acceptance. *Accid Anal Prev, 41*(2), 276-285. https://doi.org/10.1016/j.aap.2008.11.006

Chiu, S. W., Gervan, S., Fairbrother, C., Johnson, L. L., Owen-Anderson, A. F. H., Bradley, S. J., & Zucker, K. J. (2006). Sex-Dimorphic Color Preference in Children with Gender Identity Disorder: A Comparison to Clinical and Community Controls. *Sex Roles, 55*(5-6), 385-395. https://doi.org/10.1007/s11199-006-9089-9

David Leonard, S. (1999). Does color of warnings affect risk perception?.pdf. *International Journal of Industrial Ergonomics, 23*, 499-504. https://doi.org/10.1016/S0169-8141(98)00015-8

Dzulkifli, M. A., & Mustafar, M. F. (2013). The Influence of Colour on Memory. *The Malaysian journal of medical sciences: MJMS, 20,2*.

Elliot, A. J., & Maier, M. A. (2007). Color and Psychological Functioning. *Current Directions in Psychological*

Science, 16(5), 250-254. https://doi.org/10.1111/j.1467-8721.2007.00514.x

Eysenbach, G. (2020). How to Fight an Infodemic: The Four Pillars of Infodemic Management. *Journal of Medical Internet Research, 22*(6), e21820. https://doi.org/10.2196/21820

Grinstein, A., & Kronrod, A. (2016). Does Sparing the Rod Spoil the Child? How Praising, Scolding, and an Assertive Tone Can Encourage Desired Behaviors. *Journal of Marketing Research, 53*(3), 433-441. https://doi.org/10.1509/jmr.14.0224

Halpern, D., Valenzuela, S., Katz, J., & Miranda, J. P. (2019). From Belief in Conspiracy Theories to Trust in Others: Which Factors Influence Exposure, Believing and Sharing Fake News. In (pp. 217-232). Springer International Publishing. https://doi.org/10.1007/978-3-030-21902-4_16

Hastings, G., Stead, M., & Webb, J. (2004). Fear appeals in social marketing: Strategic and ethical reasons for concern. *Psychology and Marketing, 21*(11), 961-986. https://doi.org/10.1002/mar.20043

Hochberg, I., Allon, R., & Yom-Tov, E. (2020). Assessment of the Frequency of Online Searches for Symptoms Before Diagnosis: Analysis of Archival Data. *Journal of Medical Internet Research, 22*(3), e15065. https://doi.org/10.2196/15065

Kaya, N., & Epps, H. H. (2004). Relationship between color and emotion: A study of college students. *College student journal, 38*, 396.

Kim, K., & Kim, H.-S. (2018). Time Matters: Framing Antismoking Messages Using Current Smokers' Preexisting Perceptions of Temporal Distance to Smoking-Related Health Risks. *Health Communication, 33*(3), 338-348. https://doi.org/10.1080/10410236.2016.1266579

Kline, P. B., Braun, C. C., Peterson, N., & Silver, N. C. (1993). The Impact of Color on Warnings Research. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 37*(14), 940-944. https://doi.org/10.1177/154193129303701402

Klohn, L. S., & Rogers, R. W. (1991). Dimensions of the severity of a health threat The persuasive effects of visibility, time of onset, and rate of onset on young women's intentions to prevent osteoporosis. *Health Psychology, 10(5)*, 323–329. https://doi.org/https://doi.org/10.1037/0278-6133.10.5.323

Kordzadeh, N., & Warren, J. (2014, 2014). Communicating Personal Health Information in Virtual Health Communities: A Theoretical Framework. 2014 47th Hawaii International Conference on System Sciences, (pp. 636-645). IEEE.

Kordzadeh, N., & Warren, J. (2017). Communicating Personal Health Information in Virtual Health Communities: An Integration of Privacy Calculus Model and Affective Commitment. *Journal of the Association for Information Systems, 18*(1), 45-81.

Kronrod, A., Grinstein, A., & Wathieu, L. (2012). Enjoy! Hedonic Consumption and Compliance with Assertive Messages. *Journal of Consumer Research, 39*(1), 51-61. https://doi.org/10.1086/661933

Kurt, S., & Osueke, K. K. (2014). The Effects of Color on the Moods of College Students. *SAGE Open, 4*(1), 215824401452542. https://doi.org/10.1177/2158244014525423

Laroche, M., Toffoli, R., Zhang, Q., & Pons, F. (2015). A cross-cultural study of the persuasive effect of fear appeal messages in cigarette advertising: China and Canada. *International Journal of Advertising, 20*(3), 297-317. https://doi.org/10.1080/02650487.2001.11104895

Lennon, R., & Rentfro, R. (2017). Are young adults fear appeal effectiveness ratings explained by fear arousal, perceived threat and perceived efficacy?

Lewis, I. M., Watson, B., White, K. M., & Tay, R. (2007). Promoting Public Health Messages: Should We Move Beyond Fear-Evoking Appeals in Road Safety? *Qualitative Health Research, 17*(1), 61-74. https://doi.org/10.1177/1049732306296395

Liberman, N., Trope, Y., & Wakslak, C. (2007). Construal Level Theory and Consumer Behavior. *Journal of Consumer Psychology, 17*(2), 113-117. https://doi.org/10.1016/s1057-7408(07)70017-7

Lupi, V. (2020). *COVID-19 AND FAKE NEWS IN THE SOCIAL MEDIA*. https://www.fbk.eu/en/press-releases/covid-19-and-fake-news-in-the-social-media/

Mady, A. (2017). Behavioral Approach to Information Security Policy Compliance Full Paper.

Meadows, C. Z. (2020). The Effects of Fear Appeals and Message Format on Promoting Skin Cancer Prevention Behaviors among College Students. *Societies, 10*(1). https://doi.org/10.3390/soc10010021

Nan, X. (2007). Social Distance, Framing, and Judgment: A Construal Level Perspective. *Human Communication Research, 33*(4), 489-514. https://doi.org/10.1111/j.1468-2958.2007.00309.x

Omura, M., Maguire, J., Levett-Jones, T., & Stone, T. E. (2017). The effectiveness of assertiveness communication training programs for healthcare professionals and students: A systematic review. *International Journal of Nursing Studies, 76*, 120-128. https://doi.org/10.1016/j.ijnurstu.2017.09.001

Ortutay, B. (2021). *Twitter Rolls Out Redesigned Misinformation Warning Labels*. abc News. Retrieved April 14, 2022 from https://abcnews.go.com/Technology/wireStory/twitter-rolls-redesigned-misinformation-warning-labels-81212172

Pipaş, M. D., & Jaradat, M. (2010). Assertive Communication Skills. *Annales Universitatis Apulensis Series Oeconomica, 2*(12), 649-656. https://doi.org/10.29302/oeconomica.2010.12.2.17

Rhodes, N. (2015). Fear-Appeal Messages: Message Processing and Affective Attitudes. *Communication Research, 44*(7), 952-975. https://doi.org/10.1177/0093650214565916

Rosch, E. (1975). Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology: General, 104(3)*, 192-233. https://doi.org/https://psycnet.apa.org/doi/10.1037/0096-3445.104.3.192

Ruiter, R. A. C., Kessels, L. T. E., Peters, G. J. Y., & Kok, G. (2014). Sixty years of fear appeal research: Current state of the evidence. *International Journal of Psychology, 49*(2), 63-70. https://doi.org/10.1002/ijop.12042

Schuetz, S. W., Benjamin Lowry, P., Pienta, D. A., & Bennett Thatcher, J. (2020). The Effectiveness of Abstract Versus Concrete Fear Appeals in Information Security. *Journal of Management Information Systems, 37*(3), 723-757. https://doi.org/10.1080/07421222.2020.1790187

Schuetz, S. W., Lowry, P. B., Pienta, D. A., & Thatcher, J. B. (2021). Improving the Design of Information Security Messages by Leveraging the Effects of Temporal Distance and Argument Nature. *Information Systems & Economics eJournal*.

Schuetz, S. W., Lowry, P. B., & Thatcher, J. B. (2016). Defending Against Spear-Phishing Motivating Users Through Fear Appeal Manipulations.pdf. *20th Pacific Asia Conference on Information Systems (PACIS 2016)*. https://doi.org/https://ssrn.com/abstract=2861410

Sharevski, F., Alsaadi, R., Jachim, P., & Pieroni, E. (2021). Misinformation Warning Labels: Twitter's Soft Moderation Effects on COVID-19 Vaccine Belief Echoes. https://doi.org/10.48550/arxiv.2104.00779

Silic, M., & Cyr, D. (2016). Colour Arousal Effect on Users' Decision-Making Processes in the Warning Message Context. In *HCI in Business, Government, and Organizations: Information Systems* (pp. 99-109). https://doi.org/10.1007/978-3-319-39399-5_10

Silic, M., Silic, D., & Oblakovic, G. (2016). The Effects of colour on users compliance with warning banner messages across cultures.pdf.

Singh, S. (2006). Impact of color on marketing. *Management Decision, 44*(6), 783-789. https://doi.org/10.1108/00251740610673332

Stewart, D. W., & Martin, I. M. (1994). Intended and Unintended Consequences of Warning Messages: A Review and Synthesis of Empirical Research. *Journal of Public Policy & Marketing, 13*(1), 1-19. https://doi.org/10.1177/074391569401300101

Sylvia Chou, W.-Y., & Gaysynsky, A. (2020). A Prologue to the Special Issue: Health Misinformation on Social Media. *American Journal of Public Health, 110*(S3), S270-S272. https://doi.org/10.2105/ajph.2020.305943

Sylvia Chou, W.-Y., Gaysynsky, A., & Cappella, J. N. (2020). Where We Go From Here: Health Misinformation on Social Media. *American Journal of Public Health, 110*(S3), S273-S275. https://doi.org/10.2105/ajph.2020.305905

Tannenbaum, M. B., Hepler, J., Zimmerman, R. S., Saul, L., Jacobs, S., Wilson, K., & Albarracin, D. (2015, Nov). Appealing to fear: A meta-analysis of fear appeal effectiveness and theories. *Psychol Bull, 141*(6), 1178-1204. https://doi.org/10.1037/a0039729

Torres, A., Serra, J., Llopis, J., & Delcampo, A. (2020). Color preference cool versus warm in nursing homes depends on the expected activity for interior spaces. *Frontiers of Architectural Research, 9*(4), 739-750. https://doi.org/10.1016/j.foar.2020.06.002

Trope, Y., & Liberman, N. (2010). Construal-level theory of psychological distance. *Psychological Review, 117*(2), 440-463. https://doi.org/10.1037/a0018963

Wang, C. X., & Zhang, J. (2020). Assertive Ads for Want or Should? It Depends on Consumers' Power. *Journal of Consumer Psychology, 30*(3), 466-485. https://doi.org/10.1002/jcpy.1165

Wiesenfeld, B. M., Reyt, J.-N., Brockner, J., & Trope, Y. (2017). Construal Level Theory in Organizational Research. *Annual Review of Organizational Psychology and Organizational Behavior, 4*(1), 367-400. https://doi.org/10.1146/annurev-orgpsych-032516-113115

Witte, K. (1992). Putting the fear back into fear appeals: The extended parallel process model. *Communication Monographs, 54*, 329-349. https://doi.org/10.1080/03637759209376276

Witte, K., & Allen, M. (2000). A Meta-Analysis of Fear Appeals: Implications for Effective Public Health Campaigns. *Health Education & Behavior, 27*(5), 591-615. https://doi.org/10.1177/109019810002700506

Zaaba, Z. F., Lim Xin Yi, C., Amran, A., & Omar, M. A. (2021). Harnessing the Challenges and Solutions to Improve Security Warnings: A Review. *Sensors, 21*(21), 7313. https://doi.org/10.3390/s21217313

Zhang, X., & Zhou, S. (2020). Sharing health risk messages on social media: Effects of fear appeal message and image promotion. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace, 14*(2). https://doi.org/10.5817/cp2020-2-4