# Trust and Transparency: An Exploratory Study on Emerging Adults' Interpretations of Credibility Indicators on Social Media Platforms

Erica Shusas
ejr93@drexel.edu
Drexel University
Philadelphia, PA, USA

Andrea Forte
fortea@umich.edu
University of Michigan
Ann Arbor, MI, USA

## ABSTRACT

The misinformation crisis across social media has disrupted critical access to information in health, politics, and public safety. Content labels have become a feature that social media platforms use to signal credibility of social media posts. Young adults receive a proportionally high percentage of their news through social media platforms, yet prior work has shown that credibility indicators are not effective signals for young audiences. This late-breaking work presents initial findings from an exploratory study into how emerging adults (ages 18-25) assess different credibility indicators currently used on social media platforms. Our findings indicate that participants have a wide variety of interpretations of the purpose and source of context labels, are supportive of automated approaches to content labeling, and trust social media platforms to oversee the application of content labels. This paper contributes these findings to the growing scholarship on content labeling and discusses their implications for designers and policymakers.

## CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; **Empirical studies in HCI**.

## KEYWORDS

credibility indicators, emerging adults, credibility assessment, young people, misinformation

**ACM Reference Format:**
Erica Shusas and Andrea Forte. 2024. Trust and Transparency: An Exploratory Study on Emerging Adults' Interpretations of Credibility Indicators on Social Media Platforms. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3613905.3650801

## 1 INTRODUCTION

In the last few years, we have seen the detrimental effects of misinformation disrupting people's ability to access reliable information related to topics such as health [11], elections [13], wars [23], and politics [20]. Partly due to the ease of sharing content, social media platforms have become a breeding ground for the spread of misinformation [26]. Many approaches have been explored to inform social media users that an article has been identified as misinformation through various credibility indicators, including warning messages from fact-checkers [14], context labels [38], and the display of social endorsement cues. While some of these approaches have shown promise [16], the search for trusted and effective misinformation labeling is ongoing and complex. Automated approaches have shown encouraging results in accurately detecting false information and could help ease the speed with which it spreads, but studies have shown that people's trust in automated approaches lags behind human fact-checkers [22, 37]. The public has also been divided in their level of trust in social media platforms as gatekeepers of credible content indicators. Public attitudes toward the role that social media companies should have in moderating content have been divided between those who believe technology companies should take action to restrict misinformation online and those who feel that social media companies should prioritize freedom of speech [18]. Moreover, individuals from different demographic groups, such as age [15, 37], political ideology [12], and education level [29], have been shown to have different responses to credibility indicators.

One of these demographic groups that is of particular interest is young adults. Although they depend on social media platforms for a proportionally high percentage of their news [30], research has shown that younger age groups are more resistant to credibility indicators [15, 37]. How do we design credibility indicators that are both trustworthy and effective with young people? As a step toward answering this question, we designed an exploratory study to investigate how emerging adults interpret credibility indicators in use today on social media platforms. We performed semi-structured interviews along with a think-aloud protocol with 35 individuals between the ages of 18-25 to investigate how they would assess the credibility of six different social media posts, three with a credibility indicator (in the form of a context label) and three without. Our initial findings indicate that participants had a wide variety of interpretations of the purpose and source of the context labels, are largely supportive of automated approaches to content labeling, and trust social media platforms to oversee the application of content labels. In this late-breaking work, we discuss these findings and their implications for designers and policymakers, along with plans for future work.


*CHI EA '24, May 11–16, 2024, Honolulu, HI, USA*

ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3650801

# 2 RELATED WORK

## 2.1 Young people and credibility assessment

In part due to their relatively high level of engagement with online information resources compared with other age demographics, the way young people evaluate online information has been a research focus across multiple academic disciplines for over twenty years [7, 9, 17, 21]. Recent work at CHI has been influential to scholarship in this area. For example, in a mixed methods study with 35 participants from Generation Z, Hassoun et al. [10] found that members of this age group are shaped more by social motivations than by truth-seeking queries. Additionally, using data from a survey and co-design study of young people's online help-seeking information behavior, Pretorius et al. [24] proposed four design recommendations (connectedness, accessible information, personalization, and immediacy) to address the challenges of young adulthood, including transitioning into independent living, body image issues, and education. While prior work has looked at the effects that misinformation on social media platforms has on younger age groups [2, 4] and how young people share misinformation [1, 3], there is a dearth of research on how young people interpret the credibility indicators that are in use on social media platforms to help identify misinformation.

## 2.2 Credibility indicators

Credibility indicators, such as fact-check labels, click-through barriers that include a warning, content-sensitivity alerts, and context labels that provide additional information, are tools that social media platforms use to moderate problematic content [19]. Prior work has investigated the effectiveness of different credibility indicators to help individuals identify misinformation [6, 33, 35, 37]. Recent studies have shown that people find human fact-checkers more trustworthy compared with automated approaches [22, 37]. However, other work has revealed less variance among fact-checking sources [15] and has shown that demographic factors, such as political ideology [12], can impact the level of effectiveness of fact-checker, community, and algorithmic labels. Research has also looked at how credibility indicators impact different age groups in their propensity to share misinformation and their perceived accuracy of content. Yaqub et al. [37] conducted a study on the effects of four types of sources of credibility dispute (fact-checking journalists, major news outlets, a majority of Americans, and algorithms using Artificial Intelligence (AI) techniques) on people's intent to share news headlines with friends on social media and found that the credibility indicators were less impactful on young people (ages 18-29), who more frequently reported intentions to share non-true headlines. Additionally, Liu et al. [15] found that fact checks were more effective at reducing belief in misinformation in older demographics. Guo et al. [8] performed an interview and think-aloud study with 28 18-25-year-olds on how their interactions with one interstitial and two contextual warning labels (one general and one specific) influenced their perceived accuracy of short video content. They found that specific contextual warning labels do not always evoke behavioral adherence but can warn users about misinformation and that general contextual warning labels were easily disregarded due to their ambiguity. We contribute to this scholarship by offering insight into how emerging adults interpret context labels in use on social media platforms and how they align with their credibility assessment strategies.

## 2.3 Content label application

Another consideration in content labeling is who should be trusted to apply the label itself. In the United States, content labeling has historically either been self-imposed or mandated by government regulation [34]. American social media companies' role in stemming the misinformation spread on their platforms has been hotly contested, and content labels have emerged as a middle ground between a laissez-faire tactic and more punitive approaches like censoring or downranking posts [36]. In the recent past, public attitudes toward government regulation of online content have been mixed. Pew reported in 2021 that about half of U.S. adults (48%) felt the government should take measures to restrict false online information, while the other half (50%) believe freedom of information should be protected, even at the expense of publishing some misinformation (down from 58% in 2018). Consistent with reports from 2018, 59% of Americans believed technology companies should take steps to restrict misinformation online, compared with 39% who felt that social media companies should prioritize freedom of speech [18]. Recent work at CHI reflects some of these conflicting views. For example, in a co-design and diary study, Saltz et al. [27] found a strong divide between participants who felt that social media platforms have a responsibility to apply labels to content they know is false and participants who believed that platforms should not be trusted to label content because they are politically motivated and biased. We build upon this work to explore emerging adults' attitudes toward the role social media platforms should have in labeling content.

# 3 METHOD

Participants were recruited through convenience sampling. Flyers were distributed around the campus and adjoining neighborhoods of a U.S. urban university campus and posted on the first author's social media accounts. The study requirements were for participants to be between the ages of 18 and 25 and social media users. Potential participants were asked to self-report their eligibility before being accepted into the study and self-report their demographic information at the conclusion of the study. Participants were provided with a study information sheet, and verbal consent was obtained. Interviews lasted about an hour, and participants were compensated with a $25 gift card. All interviews took place between May and December 2022, with approval from our institution's IRB. Our final participant group included 35 people between the ages of 18 and 25 (9 female, 26 male; 32 Black/African American, 2 White, 1 Black/Native; 4 Democrat, 2 Republican, 1 Republican/Independent, 3 Independent; and 24 with no political affiliation).

All interviews were conducted over Zoom. In the first part of the interview, the participants were asked about their experiences with finding information on social media sites and how they determine whether a piece of online information is true. The second part consisted of a think-aloud protocol. Participants were asked to verbalize their thoughts and what they would do when assessing the validity of six different social media posts from three different platforms–two posts from each platform, three with context labels

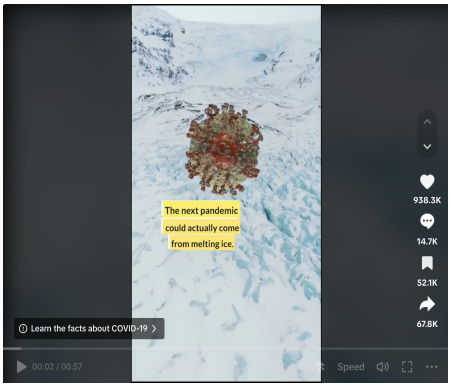


**Figure 1: Label a**

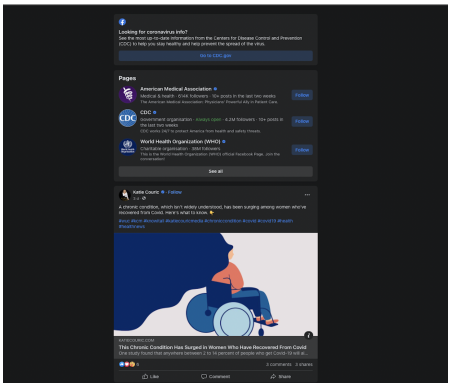


**Figure 2: Label b**

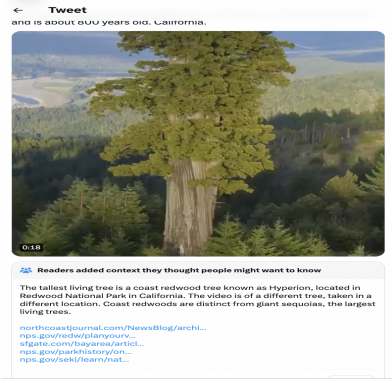


**Figure 3: Label c**

and three without–visible to them through a screen share. Although the three labels could all be considered context labels, each label had a few distinct characteristics that we felt might spark discussion. The first (Figure 1) was a TikTok context label stating "Learn the facts about COVID-19" that linked to a separate page on TikTok that lists community resources related to COVID and TikTok's medical misinformation policies. The second (Figure 2) was a Facebook label stating "Looking for coronavirus info?" followed by a short explanation that the link would take you to CDC.gov for up-to-date information on the virus. The third (Figure 3) was an X (Twitter) label with the header "Readers added context they thought people might want to know" from an X feature called Community Notes that allows X users to add a note with additional context and/or links intended to help other X users evaluate the credibility of the post. In this study, we were not concerned with whether the participant correctly identified the accuracy of the post; we focused on their interpretation of the label and how they situated it within their credibility assessment of the post. The interview guide is available in the supplementary materials.

We performed an inductive, qualitative analysis to understand the themes present in the data. The audio recordings were transcribed using Rev.com. The authors met regularly to discuss emergent themes and to develop a codebook. We used the constant comparison method by continually comparing and updating codes and themes. [5].

## 4 FINDINGS

Participants shared a wide range of credibility assessment strategies that we report in Figure 4. This figure represents credibility assessment strategies that participants reported using in the past or in relation to at least one of the six social media posts they evaluated during their interview. We focus on our three main observations in the sections below. To distinguish when a participant is referring to a particular label, we note it following a colon (e.g., P1:a, for Participant 1, label a).

### 4.1 Purpose and source of label

Participants reported many varying explanations for both the purpose and the source of all three of the context labels we discussed during our interviews. Among these explanations, we heard that the label's purpose was to direct you to other related posts on the platform (P9:a, P13:a, P17:a, P20:a, P22:a, P32:a, P32:b, P33:b), to offer personalized recommendations based on the user's profile or activity (P4:a, P6:b, P8:a, P8:b, P12:b, P21:b, P22:b, P24:a, P27:b), and to promote the individual or organization the label was linked to (P24:c, P26:a, P26:c). Most participants reported that at least one purpose of the labels was to offer more information (P1-P3, P5-P15, P17-P18, P20-P21, P24, P26-P34), but not necessarily to help verify the content. For example, P7 explained:

> "I might use it to learn more about, um, about what's going on in the video, uh, but that doesn't, um, you know, verify the credibility of the information." (*P7:c*)

Eight times we heard that the labels were there to help people verify information (P13:b, P19:c, P21:c, P23:c, P27:c, P31:c, P32:c).

We frequently heard participants report that they assumed the labels were added by the poster of the content (P3:c, P4:a, P11:a, P11:b, P12:a, P14:a, P15:c, P16:c, P18:a, P19:c, P21:a, P21:c, P23:a, P23:c, P24:c, P26:a, P31:a, P31:a, P32:c, P33:a). The perception that the label was provided by the person who posted the content often roused suspicion. For example, P10 explained his suspicion of label c, which he perceived to have been created by the poster of the content:

> "Yeah, since it's an individual's and the post which has been put there is also from an individual, so the information may be true or may be false." (*P19:c*)

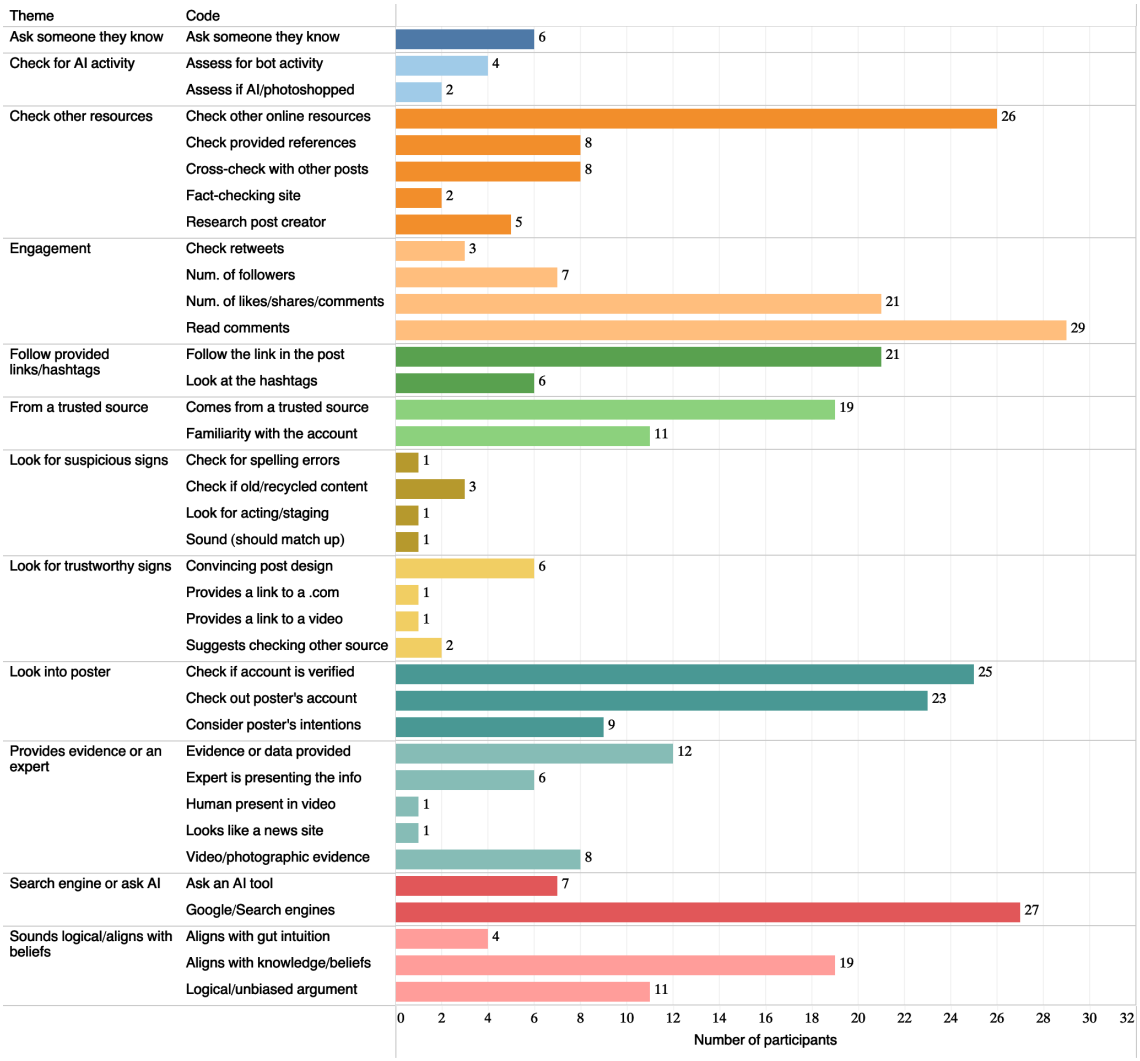


**Figure 4: Number of participants that reported using a credibility assessment strategy in the past or in relation to at least one of the six social media posts they evaluated during their interview**

Although participants frequently reported a desire to use provided links to evaluate a post, the uncertainty about where the link was taking them led several participants to share their concerns about the labels potentially being a scam or spam:

> "I don't just click on, on things on websites because I'm really, um, really, really skeptical about that because, you know, you could, you know, be malware or just the virus. So I don't click on links, but I wouldn't click on that. I would be really skeptical about it." (*P15:a*)

Overall, we found that the lack of clarity surrounding the purpose and source of context labels often led to a lack of trust in using them for information verification.

## 4.2 Trust in automated approaches

We found that most participants had confidence in automated approaches to apply contextual labels, which is not aligned with prior studies showing that people mistrust automated approaches to flagging misinformation [28, 32]. Many of the participants mentioned a strength of automated approaches is their ability to check large amounts of content in a relatively short period of time (P15, P16, P17, P19, P32). Additionally, a few of the participants (P15, P24) underscored how recent advancements in AI have improved the accuracy of automated approaches. For example, P15 explained;

> "Um, in this day and age, we have seen what AI can do, you know, of a lot of, um, AI, um, powered inventions and innovations...So I, I believe if you can create an algorithm, an algorithm and an AI that can do that for your things better than people doing it." (*P15*)

Participants also reported that they preferred the ability of automated approaches to consult several different viewpoints rather than depending on the judgment of one source (P15, P16, P23). P16 explained that:

> "I mean, it's, it has, um, access to loads and loads, loads of articles, it's in the systems. So, um, so most of the information is, um, digitalized and it's in the system, so the AI can assess hundreds and thousands of articles and, um, information per minute. And we, as human beings, we don't have that, you know, kind of time... " *(P16)*

Some participants said they trust automated approaches because they reflect knowledge of the people who program them (P6, P20). P20 stated that she'd trust an AI or a computer program because:

> "Their sources of information are credible and even the information provided because the people [programming] the information have expertise on the topics and they're mainly researchers, data analysts." *(P20)*

Overall, participants didn't mention many disadvantages of automated approaches. A few exceptions included P33, who shared that, in general, she felt that AI does more harm than good, and P34, who thought the labels he saw that appeared to be automated were too "generalized" to be helpful. It was more common for participants to report that automated approaches added a valuable, if not comprehensive, contribution to credibility assessment on social media (P6, P8, P10, P13, P14, P18, P22, P24).

## 4.3 The role of social media platforms

Prior work has revealed deep divides in people's opinions about the responsibility and authority social media platforms should have to label content [27]. However, we found that participants were largely supportive of social media platforms' involvement in misinformation detection and labeling. Sometimes, participants shared that they believed that social media platforms were motivated to try to help people detect trustworthy information so that they would not harm their business reputation (P7, P8, P12). P12 explained that:

> "Yeah, if a group of people at Twitter were putting the, I would, um, believe it more because, you know...they don't want to tarnish the reputation of Twitter, so they would've to, you know, verify information before posting it." (*P12:c*)

Some participants shared a sense that social media platforms have benevolent intentions when delivering context labels. As P8 put it:

> "Yeah, I would like, I would click on, because most of these social media platforms, I know they have a way like to try to help people mm-hmm. <affirmative>, let's say if you go to a social media website and search something like, I want to kill myself. So most probably they will attach you a link trying to tell you how you can solve your problems or to talk to somebody. So I would, uh, I would click a link like this and, uh, see the information that is in need." (*P8*)

Participants also revealed an expectation that social media platforms were overseeing the moderation of content (P2, P7). For example, P2 shared that he would trust community notes because they were probably under the watch of Twitter developers:

> "It's Twitter based, so I trust it. And, uh, uh, yeah, basically they shed light on information that's, that's through Twitter to differentiate between what's real and what's not real." (*P2*)

## 4.4 Design recommendations

At the end of the interview, participants were asked if there were any features they would like to see social media platforms implement to help them determine if a post was true. The most common response involved the placement of reliable source links alongside posts (6 participants). Other responses included the ability to up or down vote posts (2 participants), more moderation and tagging of false posts by a moderator (1 participant) or algorithm (1 participant), more fact-checked labels (1 participant), a reliability score for posts (1 participant), each poster has the history of the reliability of their past posts accessible (1 participant), and repeat posters of misinformation should receive consequences (1 participant). Markedly, most of the recommendations involved the inclusion of reliable source links and additional content labels.

# 5 DISCUSSION

In this exploratory study, we examined how emerging adults interpret credibility indicators in use on social media platforms. Our findings contribute valuable insights for social media platform designers and policymakers. We discuss the implications of each of our primary observations below.

Although our sample is limited, we had a surprisingly wide diversity of explanations for both the purpose and the source of all three of the context labels we discussed during our interviews. Sometimes, this uncertainty caused an unwillingness to engage with the label. Our study participants spoke of their experiences and awareness of the harms that can be caused by misinformation and scams. Uncertainty about the source of the context label at times led them to believe that it may lead them to the same type of problematic content that the label is intended to prevent. This finding suggests to platform designers that clearer explanations of the purpose and source of labels need to be communicated with users in order for them to be effective.

Our study's participants largely trusted automated approaches to context labels, which is not aligned with prior work [22, 37]. Seven of our participants also reported consulting an AI tool as one of their credibility assessment strategies. We postulate that perhaps recent technological advancements and growth in usage of chatbots for information retrieval by young people [31] may have heightened their level of trust in algorithmic and AI-powered content labeling. Although the question of if and when one should trust an AI tool to verify information is beyond the scope of this study, the growing usage of AI tools by younger age demographics might engender a growing trust in automated approaches to content flagging within this age group.

Finally, we observed an unexpectedly high level trust in social platforms to monitor and deploy dependable content labeling tools.

Reading comments and looking at social endorsement cues, such as likes and shares, was a common credibility assessment strategy mentioned by participants (Figure 4). However, participants also often spoke of the tendency of many of their peers to post content for the purpose of likes and attention, which led them to view posts with more caution. When discussing label c, the community note, we observed participants finding value in links posted by other users as a way to share opinions, but they often weren't seen as a means to verify the post. Although we found that participants wanted to be able to observe the level of engagement a post has, we also heard a desire for more moderation from the platform. This observation is interesting in light of a recent UNICEF-Gallup poll [25] that found that while young people between the ages of 15 and 24 rely on social media to stay informed about current events more than any other age demographic (45%), only 17% put their trust in social media content. According to the poll, young people were also found to be more trusting in institutions (other than the police) than older demographics, including government, scientists, medical professionals, and national and international news organizations. While social media platforms were not included in this study, our study participants frequently shared a preference for and a trust in platform-moderated content labeling over labels they perceived to be provided from other users. This finding is valuable when designing platforms for younger demographics and may also warrant further exploration by policymakers when crafting social media content moderation regulations.

## 5.1 Limitations, Future work, and Conclusion

We present this interview and think-aloud study of 35 individuals between the ages of 18 and 25 as a step toward understanding how to build better credibility indicators on social media platforms for emerging adults. Our sample is limited and does not represent the U.S. or global emerging adult population. Additionally, there are many other types of credibility indicators in use on social media platforms that we did not include in this study, and there are more differences to explore between the three labels we did include. However, our findings contribute insights that can have profound implications for platform designers and policymakers. First, we found that participants have a wide variety of perceptions of the purpose and source of contextual content labels and desire more transparency with both of these features. Second, we found participants to be largely trusting of automated approaches to content labeling. Third, our participants were trusting of social platforms to monitor and deploy reliable content labeling tools. Our next step in this study is to deploy a survey that is informed by the results of this qualitative inquiry that will dig deeper into these initial findings and can better understand how demographic factors within the emerging adult population may affect perceptions and attitudes toward credibility indicators on social media platforms.

## REFERENCES


[1] Ifeoma Adaji. 2023. Age Differences in the Spread of Misinformation Online. *European Conference on Social Media* 10, 1 (May 2023), 12–19. https://doi.org/10.34190/ecsm.10.1.1156

[2] Dolores Albarracin, Daniel Romer, Christopher Jones, Kathleen Hall Jamieson, and Patrick Jamieson. 2018. Misleading Claims About Tobacco Products in YouTube Videos: Experimental Effects of Misinformation on Unhealthy Attitudes. *Journal of Medical Internet Research* 20, 6 (June 2018), e229. https://doi.org/10.2196/jmir.9959

[3] Vimala Balakrishnan. 2022. Socio-demographic Predictors for Misinformation Sharing and Authenticating amidst the COVID-19 Pandemic among Malaysian Young Adults. *Information Development* (Aug. 2022), 026666692211189. https://doi.org/10.1177/02666669221118922

[4] Porismita Borah, Bimbisar Irom, and Ying Chia Hsu. 2022. 'It infuriates me': examining young adults' reactions to and recommendations to fight misinformation about COVID-19. *Journal of Youth Studies* 25, 10 (Nov. 2022), 1411–1431. https://doi.org/10.1080/13676261.2021.1965108

[5] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis.* sage.

[6] Diego Esteves, Aniketh Janardhan Reddy, Piyush Chawla, and Jens Lehmann. 2018. Belittling the Source: Trustworthiness Indicators to Obfuscate Fake News on the Web. (2018). https://doi.org/10.48550/ARXIV.1809.00494 Publisher: arXiv Version Number: 1.

[7] Andrew J Flanagin and Miriam J Metzger. 2008. *Digital media and youth: Unparalleled opportunity and unprecedented responsibility.* MacArthur Foundation Digital Media and Learning Initiative Cambridge, MA, USA.

[8] Chen Guo, Nan Zheng, and Chengqi (John) Guo. 2023. Seeing is Not Believing: A Nuanced View of Misinformation Warning Efficacy on Video-Sharing Social Media Platforms. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Sept. 2023), 1–35. https://doi.org/10.1145/3610085

[9] Eszter Hargittai, Lindsay Fullerton, Ericka Menchen-Trevino, and Kristin Yates Thomas. 2010. Trust online: Young adults' evaluation of web content. *International journal of communication* 4 (2010), 27.

[10] Amelia Hassoun, Ian Beacock, Sunny Consolvo, Beth Goldberg, Patrick Gage Kelley, and Daniel M. Russell. 2023. Practicing Information Sensibility: How Gen Z Engages with Online Information. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–17. https://doi.org/10.1145/3544548.3581328

[11] Hendrik Heuer and Elena Leah Glassman. 2022. A Comparative Evaluation of Interventions Against Misinformation: Augmenting the WHO Checklist. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–21. https://doi.org/10.1145/3491102.3517717

[12] Chenyan Jia, Alexander Boltz, Angie Zhang, Anqing Chen, and Min Kyung Lee. 2022. Understanding Effects of Algorithmic vs. Community Label on Perceived Accuracy of Hyper-partisan Misinformation. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (Nov. 2022), 1–27. https://doi.org/10.1145/3555096

[13] Prerna Juneja, Md Momen Bhuiyan, and Tanushree Mitra. 2023. Assessing enactment of content regulation policies: A post hoc crowd-sourced audit of election misinformation on YouTube. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, Hamburg Germany, 1–22. https://doi.org/10.1145/3544548.3580846

[14] Jan Kirchner and Christian Reuter. 2020. Countering fake news: A comparison of possible solutions regarding user acceptance and effectiveness. *Proceedings of the ACM on Human-computer Interaction* 4, CSCW2 (2020), 1–27. ISBN: 2573-0142 Publisher: ACM New York, NY, USA.

[15] Xingyu Liu, Li Qi, Laurent Wang, and Miriam J. Metzger. 2023. Checking the Fact-Checkers: The Role of Source Type, Perceived Credibility, and Individual Differences in Fact-Checking Effectiveness. *Communication Research* (Oct. 2023), 00936502231206419. https://doi.org/10.1177/00936502231206419

[16] Cameron Martel and David G. Rand. 2023. Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology* (Oct. 2023), 101710. https://doi.org/10.1016/j.copsyc.2023.101710

[17] Ericka Menchen-Trevino and Eszter Hargittai. 2011. Young Adults' Credibility Assessment of Wikipedia. *Information, Communication & Society* 14, 1 (2011), 24–51. Publisher: Taylor & Francis.

[18] Amy Mitchell and Mason Walker. 2021. More Americans now say government should take steps to restrict false information online than in 2018. https://www.pewresearch.org/short-reads/2021/08/18/more-americans-now-say-government-should-take-steps-to-restrict-false-information-online-than-in-2018/

[19] Garrett Morrow, Briony Swire-Thompson, Jessica Montgomery Polny, Matthew Kopec, and John P. Wihbey. 2022. The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology* 73, 10 (Oct. 2022), 1365–1386. https://doi.org/10.1002/asi.24637

[20] Mohsen Mosleh, Cameron Martel, Dean Eckles, and David Rand. 2021. Perverse Downstream Consequences of Debunking: Being Corrected by Another User for Posting False Political News Increases Subsequent Sharing of Low Quality, Partisan, and Toxic Content in a Twitter Field Experiment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–13. https://doi.org/10.1145/3411764.3445642

[21] Tayo Oyedeji. 2011. Credibility perceptions of different types of weblogs among young adults. *Global Media Journal* 11, 19 (2011), N_A. Publisher: Purdue

University Calumet.

[22] Christina A. Pan, Sahil Yakhmi, Tara P. Iyer, Evan Strasnick, Amy X. Zhang, and Michael S. Bernstein. 2022. Comparing the Perceived Legitimacy of Content Moderation Processes: Contractors, Algorithms, Expert Panels, and Digital Juries. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW1 (March 2022), 1–31. https://doi.org/10.1145/3512929

[23] Francesco Pierri, Luca Luceri, Nikhil Jindal, and Emilio Ferrara. 2023. Propaganda and Misinformation on Facebook and Twitter during the Russian Invasion of Ukraine. In *Proceedings of the 15th ACM Web Science Conference 2023*. ACM, Austin TX USA, 65–74. https://doi.org/10.1145/3578503.3583597

[24] Claudette Pretorius, Darragh McCashin, Naoise Kavanagh, and David Coyle. 2020. Searching for Mental Health: A Mixed-Methods Study of Young People's Online Help-seeking. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–13. https://doi.org/10.1145/3313831.3376328

[25] Julie Ray. 2021. Young People Rely on Social-Media, but Don't Trust it. *Gallup.com* (2021).

[26] Matthew Sadiku, Tochukwu Eze, and Sarhan Musa. 2018. Fake news and misinformation. *International Journal of Advances in Scientific Research and Engineering* 4, 5 (2018), 187–190.

[27] Emily Saltz, Claire R Leibowicz, and Claire Wardle. 2021. Encounters with Visual Misinformation and Labels Across Platforms: An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, Yokohama Japan, 1–6. https://doi.org/10.1145/3411763.3451807

[28] Haeseung Seo, Aiping Xiong, and Dongwon Lee. 2019. Trust It or Not: Effects of Machine-Learning Warnings in Helping Individuals Mitigate Misinformation. In *Proceedings of the 10th ACM Conference on Web Science*. 265–274.

[29] Filipo Sharevski, Amy Devine, Emma Pieroni, and Peter Jacnim. 2022. Meaningful Context, a Red Flag, or Both? Users' Preferences for Enhanced Misinformation Warnings on Twitter. http://arxiv.org/abs/2205.01243 arXiv:2205.01243 [cs].

[30] Elisa Shearer. 2021. More than eight-in-ten Americans get news from digital devices. *Pew Research Center* 12 (2021). https://www.pewresearch.org/short-reads/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/

[31] Rohit Shewale. 2024. ChatGPT Statistics — User Demographics (January 2024). https://www.demandsage.com/chatgpt-statistics/

[32] Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond news contents: The role of social context for fake news detection. In *Proceedings of the twelfth ACM international conference on web search and data mining*. 312–320.

[33] Niraj Sitaula, Chilukuri K. Mohan, Jennifer Grygiel, Xinyi Zhou, and Reza Zafarani. 2020. Credibility-Based Fake News Detection. In *Disinformation, Misinformation, and Fake News in Social Media*, Kai Shu, Suhang Wang, Dongwon Lee, and Huan Liu (Eds.). Springer International Publishing, Cham, 163–182. https://doi.org/10.1007/978-3-030-42699-6_9 Series Title: Lecture Notes in Social Networks.

[34] Matthew Spradling, Jeremy Straub, and Jay Strong. 2021. Protection from 'Fake News': The Need for Descriptive Factual Labeling for Online Content. *Future Internet* 13, 6 (May 2021), 142. https://doi.org/10.3390/fi13060142

[35] Julian Unkel and Alexander Haas. 2017. The effects of credibility cues on the selection of search engine results. *Journal of the Association for Information Science and Technology* 68, 8 (2017), 1850–1862. ISBN: 2330-1635 Publisher: Wiley Online Library.

[36] John Wihbey, Matthew Kopec, and Ronald Sandler. 2021. Informational Quality Labeling on Social Media: In Defense of a Social Epistemology Strategy. *Available at SSRN 3858906* (2021).

[37] Waheeb Yaqub, Otari Kakhidze, Morgan L. Brockman, Nasir Memon, and Sameer Patil. 2020. Effects of Credibility Indicators on Social Media News Sharing Intent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, Honolulu HI USA, 1–14. https://doi.org/10.1145/3313831.3376213

[38] Himanshu Zade, Megan Woodruff, Erika Johnson, Mariah Stanley, Zhennan Zhou, Minh Tu Huynh, Alissa Elizabeth Acheson, Gary Hsieh, and Kate Starbird. 2023. Tweet Trajectory and AMPS-based Contextual Cues can Help Users Identify Misinformation. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (April 2023), 1–27. https://doi.org/10.1145/3579536