WARNING This Contains Misinformation: The Effect of Cognitive Factors, Beliefs, and Personality on Misinformation Warning Tag Attitudes

ROBERT KAUFMAN, University of California, San Diego, USA
AARON BROUKHIM, University of California, San Diego, USA
MICHAEL HAUPT, University of California, San Diego, USA

Social media platforms enhance the propagation of online misinformation by providing large user bases with a quick means to share content. One way to disrupt the rapid dissemination of misinformation at scale is through warning tags, which label content as potentially false or misleading. However, past warning tag mitigation studies yield mixed results for diverse audiences. We hypothesize that personalizing warning tags to the individual characteristics of their diverse users may enhance mitigation effectiveness. To reach the goal of personalization, we need to understand how people differ and how those differences predict a person's attitudes and behaviors toward tags and tagged content. In this study, we leverage Amazon Mechanical Turk (n = 132) and undergraduate students (n = 112) to provide this foundational understanding. With all participants combined, we find attitudes towards warning tags and self-described behaviors are significantly influenced by factors such as Need for Cognitive Closure (NFCC), Political orientation, and Trust in Medical Scientists when controlled for covariates such as age and recruiting platform. Analyses of each sample further show that tag attitudes were influenced by Trust in Religious Leaders, and Big Five Inventory (BFI) traits for Openness and Conscientiousness. We synthesize these results into design insights and a future research agenda for more effective and personalized warning tags and misinformation mitigation strategies more generally.

CCS Concepts: • **Human-centered computing** → **User models**; **Collaborative and social computing**.

Additional Key Words and Phrases: Misinformation, Social Media, Bias, Individual Differences

ACM Reference Format:
Robert Kaufman, Aaron Broukhim, and Michael Haupt. 2025. WARNING This Contains Misinformation: The Effect of Cognitive Factors, Beliefs, and Personality on Misinformation Warning Tag Attitudes. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW340 (November 2025), 32 pages. https://doi.org/10.1145/3757521

1 Introduction

Misinformation, or "fake news", is a long-standing and pervasive issue with many adverse implications. Affecting a wide range of domains, misinformation has been tied to lower vaccine usage during the COVID-19 pandemic [11], climate change denial [113], and election interference [32] among other harmful effects. Social media platforms uniquely contribute to the spread of misinformation due to their large user bases, under-moderation, and the network effects of virality [7, 29, 106]. Since the rise of social media – where 62% of people get their news [30] – the spread of misinformation has dramatically increased [114]. The World Health Organization (WHO) has gone so far as to declare an "infodemic" characterized by the intentional spreading of inaccurate information to undermine public health efforts [80].

Authors' Contact Information: Robert Kaufman, rokaufma@ucsd.edu, University of California, San Diego, La Jolla, California, USA; Aaron Broukhim, aabroukh@ucsd.edu, University of California, San Diego, La Jolla, California, USA; Michael Haupt, mhaupt@ucsd.edu, University of California, San Diego, La Jolla, California, USA.

ACM 2573-0142/2025/11-ARTCSCW340
https://doi.org/10.1145/3757521

As the problem of misinformation continues to grow, it is critical to develop effective mitigation methods that can work at scale. Several approaches have been taken to prevent false belief formation from misinformation as well as curb misinformation spread, each approach depending on the ubiquitousness of the case. Early algorithmic intervention – often using artificial intelligence or crowdsourcing – can be highly effective at moderating content visibility and preventing misinformation from spreading in the first place [29, 56, 60, 97, 98]. There are far fewer options for misinformation that has already become pervasive on social platforms – i.e. "gone viral". For these cases, tagging potentially misleading content with misinformation warning labels may be effective [17], but warnings must be *convincing* and *trustworthy* to the viewer in order to be taken seriously. This is a tall order, as different people perceive and act on information in different ways [37, 56, 84, 95]. If done correctly, warning tags present a potentially transparent [104], prudent [101], and effective [72] means of intervention that prevents the spread of misinformation.

In this paper, we focus on warning tags for misinformation on social media, as these are a front-line defense to prevent the proliferation of fake news at the point of human interaction and information uptake. We define "warning tags" as communicative actions that are spatially and temporally attached to a piece of information, news story, or source. The purpose of the tag is to alert the viewer that the content they are about to view is potentially unreliable, false, and/or encouraging them to exercise caution before accepting or sharing the information [17]. This is a common method for misinformation prevention on social media, and has been used on popular social media platforms such as X (formerly Twitter) and Facebook [75, 83]. Figure 1 shows an example from X shown to participants in the present study. Despite the wide adoption of tags across social media platforms, past studies show mixed results regarding the effectiveness of tags in combating misinformation belief and spread. Differences in efficacy are often attributed to political orientation and cognitive ability [67, 84, 95]; while these factors may predict tag effectiveness in some cases, we posit that additional characteristics may provide further insight into what types of tags work best, and for whom.

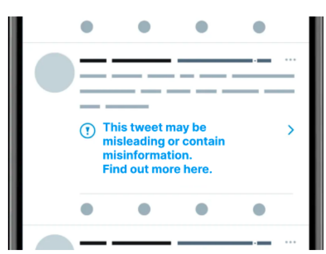

Fig. 1. Warning tag example shown to participants (from [18]). News articles were presented with the prompt above.

An intuitive method to mitigate group differences in tag efficacy is to *tailor* communications to the diverse needs of the audience. In practice, few studies have explored this possibility [101], and no major social media platform – to our knowledge – has publicly implemented personalized tags in practice. A lack of personalization limits the effectiveness of the labels due to individual differences in perception and information-related behaviors [56, 67, 84, 95]. Thus, one-size-fits-all approaches to prevent misinformation uptake and spread will fail if they don't harmonize with the specific needs and motivations of a particular individual. In the case of misinformation, tags warning of potential falsehoods need to be tailored to address the perceptions of the specific recipient, while online interventions designed to convince someone away from spreading falsehoods need to be met at the motivating source. We posit that the current one-size-fits-all implementation of misinformation tags can benefit from modifications that account for personal and contextual flexibility. Personalization and information tailoring have been topics of discussion in many other domains where people need to make sense of AI-based decisions [41, 51, 103, 115], ranging from knowledge systems [69], healthcare [53, 57], image classification [81, 108], and autonomous driving [52, 54, 55]. The need for tailored messaging is also discussed in user persona work that accounts for multiple factors such as psychological traits, situational circumstances, and demographics to produce nuanced depictions of public response towards health issues [3, 40, 42, 73].

In a similar vein, effective interventions for misinformation, especially those that are visible to users, need to account for the diversity of user reactions. However, there is little known about factors that influence user attitudes towards tags, limiting efforts to develop customized interventions. Recent work showing that warning flags can lower perceived accuracy of false news headlines demonstrates the need for user-centric models in warning tag systems [8]. As reported in Barman et al. [8], traits such as neuroticism and tendency for reflective thinking showed higher trust for warning flags with explanatory text, while political conservatism was associated with distrust towards flagging systems. However, this work does not account for factors such as the sender of the post, displays of previous engagement (eg., number of likes/shares shown to users), and may not generalize to misinformation topics other than COVID-19. Other work outlining typologies of misinformation such as satire, parody, misleading advertising, and fabricated news sources [6, 112, 118] further illustrate the need for general assessments of attitudes on tag systems. The current study fills this knowledge gap by examining attitudes about warning tag systems as a main outcome, which can guide the development of personalized tag systems across contexts. In psychology, an attitude is defined as “a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor” [24]. The present study assesses attitudes of tag warning systems, which can subsequently influence social media engagement behaviors. In order to guide the design of personalized tags, we also examine how user characteristics relevant to misinformation engagement such as political beliefs and attitudes towards institutions, thinking style, and personality traits influence favorability of tag warning systems.

The findings presented in this paper follows up on previously published work by Kaufman et al. [56], where the results showed that personal factors such as individual traits and beliefs can be used to predict misinformation assessment accuracy in tweets. In the present study, we use datasets from two diverse sample populations (undergraduates and Amazon Mechanical Turk workers) to predict attitudes towards misinformation warning tags from similar traits. We measure political beliefs, institutional trust, cognitive thinking styles, and personality traits with the purpose of highlighting how individual differences across different populations (SONA and MTurk) impact attitudes toward misinformation warning tags. The purpose of this work is to establish a foundation for personalized warning tags and provide a deeper understanding of factors that need to be accounted for to effectively mitigate misinformation uptake and propagation at the individual level.

Our results indicate that the dispositional trait Need For Cognitive Closure (NFCC), Trust in Medical Scientists, and political orientation were the strongest predictors for favorable attitudes towards tags among our overall sample. Trust in Religious Leaders and recruiting platform also showed statistically significant effects on tag attitudes. When examining differences between recruiting platforms, Big Five Inventory (BFI) personality trait Openness, political orientation, and Trust in Medical Scientists were the most influential predictors among our undergraduate sample. Among a more general population sample collected via Amazon Mechanical Turk (MTurk), we found that NFCC showed the strongest association with positive warning tag attitudes followed by BFI Conscientiousness, political orientation, Trust in Medical Scientists, and Trust in Religious Leaders. Concurrent work by Barman et al. [8] also investigate the role of age, gender, education level, political ideology, CRT, and BFI on perceived accuracy, sharing behavior, and trust in flagged content for COVID-related misinformation. Notably, this work does not test NFCC, Trust in Elected Officials, Trust in News Media, Trust in Religious Leaders, and Trust in Medical Scientists. Additionally, previous work does not measure other relevant outcomes for tag engagement such as whether the user would ignore the intervention, still read the post, research the content themselves, and how the intervention influences user perception of the post author and platform. We expand on this prior work by adding further granularity to relevant outcome measures for tag systems.

In sum, the study presented here contributes:

(1) An analysis that quantifies and ranks the effect sizes of individual differences on attitudes for warning tag systems.
(2) A comparison of two widely-used human-computer interaction research populations: Amazon Mechanical Turk (MTurk) and undergraduate study participants (SONA) in the context of misinformation warning tags.
(3) Insights and design implications for personalized and user-centric warnings tags that can more effectively mitigate the spread of misinformation among diverse populations.

Taken in full, this paper provides insight into how personal factors impact a person's attitude towards misinformation warning labels, an important and necessary first step towards developing tailored interventions that are effective not just for some, but for all.

2 Related Work

2.1 Individual Differences in Misinformation Vulnerability and Propagation

Several studies have examined misinformation vulnerability and spreading behavior. In this section, we focus on *who* is most vulnerable and likely to spread misinformation and use this as a basis to inform our study on individual differences in attitudes towards tag interventions. We list out our hypotheses based on the prior work.

Increasing reliance on social media for news contributes to the rapid dissemination of misinformation on social media [30, 114]. Since 2016, foreign agents have leveraged social media to attempt to influence the United States presidential election in favor of their preferred candidate [5], damaging voter agency and potentially swaying election outcomes against the interest of the electorate. Following this election, in which the term “fake news” became pervasive in popular culture, the COVID-19 pandemic showed the influence of misinformation via debate around vaccine efficacy. A study by Loomba et al. [71] showed a negative correlation between exposure to misinformation and intent to vaccinate against the virus, with potentially detrimental health outcomes. Importantly, this study was also one of the first to highlight individual differences in misinformation susceptibility and behavioral intention to vaccinate. They found that people of different races, religious backgrounds, and genders demonstrated different levels of susceptibility to misinformation.

In the United States, prior work has consistently highlighted political differences in misinformation vulnerability [68] and sharing behavior [38, 39]. Past studies show that people who identify as politically conservative share more misinformation and may be less accurate at identifying misinformation than those who identify as liberal [33, 56, 92]. The divide is not absolute, however: both liberals and conservatives are more likely to share misinformation that positively reflects their own ideological groups [88] and seek out information that confirms their previously held beliefs [23]. Partisanship also plays a role in attitudes toward misinformation interventions: Saltz et al. [100] found heavily-leaning liberals supported a variety of online misinformation interventions, while conservatives opposed interventions outright. Due to previously observed political differences in preferences for misinformation interventions, the present study seeks to build on this past work by examining the impact of political orientation on attitudes towards tag mitigation solutions when controlled for effects from other personal traits. In the present study, we hypothesize:

- **H1: Political conservatism will be negatively associated with favorable attitudes towards misinformation tags.**

There is a mounting archive of evidence suggesting that the informational sources a person trusts may impact their susceptibility to misinformation, including interpersonal connections, institutions, and news outlets. Intuitively, trusting experts – particularly medical experts – has been shown to decrease misinformation susceptibility, while trust in family, friends, and less reliable sources increases susceptibility [13, 56, 76]. Surprisingly, individuals with high trust in science may be more susceptible to misinformation that contains pseudo-scientific content [78]. Divergent from trust in science, Trust in Religious Leaders has been associated with belief in misinformation [87]. It's worth noting that Jasinskaja-Lahti and Jetten [46] highlights a distinction between individuals who self-categorize as religious and individuals who endorse religious worldviews, stating that endorsement of religious worldviews may be a better predictor of susceptibility to conspiracy theories. Prior work by Kaufman et al. [56] found crowdworkers with less Trust in Religious Leaders to be more effective at labeling misinformation. Trust characteristics may underlie political differences in misinformation vulnerability, sharing behavior, and attitudes [1]. It is plausible that trust in other institutional authorities such as elected officials and media outlets may also influence warning tag attitudes, as suggested by previous work showing that posts from political figures and media affiliates were often shared in COVID-related misinformation discourse [36, 38, 39]. Therefore, we seek to understand how trust in informational sources including news media, elected officials, medical scientists and religious leaders may impact the viability of warning tags for misinformation mitigation. Based on the literature, we propose the following hypotheses:

- **H2: Trust in medical scientists, elected officials, and news media will be positively associated with favorable attitudes towards misinformation tags.**
- **H3: Trust in religious leaders will be negatively associated with favorable attitudes towards misinformation tags.**

Prior studies have shown that cognitive, information assessment, and personality traits may impact misinformation vulnerability and spreading behavior. Kaufman et al. [56] found that respondents with high Cognitive Reflection Test (CRT) scores, Conscientiousness, and Trust in Medical Scientists were more aligned with experts in an information assessment task while respondents with high Need for Cognitive Closure (NFCC) and those who lean politically conservative were less aligned with experts.

Often measured by the Cognitive Reflection Test (CRT), tendency for reflective thinking has been found to be correlated with misinformation sharing and intervention attitudes. CRT performance is associated with the ability to override incorrect analytical intuitions through reflection. Pehlivanoglu et al. [82] found positive correlations between CRT score and an individual's ability to identify misinformation. Another study showed individuals with lower CRT scores were more willing to *overclaim* knowledge and had a general tendency to be more receptive to "pseudo-profound bullshit" [86]. CRT has also been shown to be positively associated with trust in medical scientists [65] and negatively associated with misinformation engagement [37] and susceptibility [85]. Since CRT is a relevant factor for evaluating information veracity and online sharing behaviors, the present study tests whether CRT is also a significant predictor for warning tag attitudes. We hypothesize:

- **H4: CRT will be positively associated with favorable attitudes towards misinformation tags.**

Other information assessment traits, such as a person's Need for Cognitive Closure (NFCC), have been shown to impact vulnerability to misinformation [9, 89]. Need for Cognitive Closure (NFCC) is a measure of one's desire for order, predictability, and discomfort with ambiguity [119]. Prior work has shown NFCC to be negatively correlated with misinformation identification accuracy [56]. Other studies have shown that individuals with high NFCC may be more likely to spread misinformation on social media networks, with the effect attributed to avoidance behavior when asked to provide evidence for beliefs [9]. The effect of NFCC in other information evaluation domains has been contested [21]. Within the context of warning tags, the presence of a tag can introduce uncertainty to users by explicitly questioning the veracity of the information presented in the post. As warning tags may impact the level of ambiguity for online information assessment, we seek to examine the influence of NFCC on warning tag attitudes. We hypothesize:

- **H5: NFCC will be positively associated with favorable attitudes towards misinformation tags.**

Prior work has established connections between personality and misinformation vulnerability [56, 63] and spreading behavior [4, 37, 43, 66]. The Big Five Inventory (BFI) Conscientiousness scale, for example, measures an individual's tendency to be orderly, cautious, and self-disciplined. Conscientiousness has been associated with better news discernment [16], and individuals low in Conscientiousness may be more likely to share fake news [66]; although, later research refutes these claims [70]. BFI Openness measures a person's curiosity, receptivity to new experiences, and imagination. Some reports find that individuals high in Openness may be more vulnerable to misinformation [2]. However, there is also conflicting evidence observed with this BFI trait: prior work demonstrates a negative association between Openness and belief in myths [111] and a positive association with news discernment [16]. BFI Agreeableness also shows conflicting evidence, as it has been associated with users who validate news prior to sharing [102] and with misinformation susceptibility [15]. Misinformation studies testing BFI traits do not often report effects from BFI Extraversion, however, there is work showing a negative association between this trait and news discernment ability [14]. BFI Neuroticism is similarly underreported in the misinformation literature, although recent work shows a positive association between Neuroticism and trust in misinformation warning tags [8]. While these traits have been used to examine misinformation susceptibility and sharing behaviors, there is little work assessing how these traits influence attitudes towards misinformation warning tags. In the present work, we assess the influence of personality traits on tag attitudes and self-report behaviors, as these dimensions may impact how information is evaluated and perceptions of interventions. We use the BFI scale to measure personality, as it is often used to evaluate the effects of individual traits on attitudes and acceptance of social networking technologies [58, 99, 110]. We hypothesize:

- **H6: BFI Conscientiousness, Openness, and Neuroticism will be positively associated with favorable attitudes towards misinformation tags.**
- **H7: BFI Agreeableness and Extraversion will be negatively associated with favorable attitudes towards misinformation tags.**

Despite all of these studies assessing the impact of individual factors on misinformation vulnerability and spread, very little work has focused on the impact of personal factors on misinformation mitigation solutions like warning tags [31] specifically. In this work, we quantify and rank the relative importance of personal factors that may influence attitudes towards misinformation tags. No prior work, to our knowledge, directly compares the impact of individual differences at the level of cognitive style, beliefs, and personality traits on general attitudes towards tag warning systems. We seek to fill these knowledge gaps.

2.2 Misinformation Warning Tags as Interventions

Due to the ubiquitous nature of misinformation spread online, deployment of mitigation solutions that can work at-scale is vital. Prior studies have found the efficacy of tags and other intervention methods to prevent belief in misinformation is complex and multi-faceted.

Attempts to curb the sharing of misinformation often involve moderation. Determining what content to moderate is a difficult problem within itself: algorithms that can work swiftly at scale typically come at the cost of nuance [29]. Recruiting a human to handle moderation edge cases can help [64], but this approach assumes underlying algorithms can identify edge cases accurately. Further, human-identified misinformation introduces the partiality of a human into the moderation process [35]. Some users may be skeptical of warning tag accuracy depending on if the correction comes from a community member or an algorithm [47] and many believe tags from humans are more biased than tags from algorithms [116]. As a result, AI-based tagging of potential misinformation and AI-generated labeling of online posts is seen as a potential soft-moderation middle ground that can help curb misinformation spread and vulnerability without completely removing the agency of the reader. Deployment efforts have been made by a number of social media platforms, including Facebook and X (formerly Twitter) [19].

While the effectiveness of warning labels for misinformation has generally seen positive results, there is an ongoing debate on their efficacy. As a whole, misinformation tags have been shown to have a modest, positive effect on reducing belief in misinformation [17, 22, 60, 83]. Initial reports of a "backfire" or "boomerang" effect that claim misinformation corrections might actually increase misperceptions [77] have since been refuted [120], however, additional work has found that topic domain may play a role in correction effectiveness [93]. For example, a meta-analysis by pui Sally Chan and Albarracín [93] found correction attempts similar to misinformation tag labels are more successful in scientific domains outside of health.

We argue that one of the reasons for prior mixed results on the efficacy of tags is due to the complex nature of misinformation vulnerability, spreading behavior, and attitudes towards mitigation efforts. Misinformation vulnerability and spreading behavior is affected by a person's psychological faculty [59], frequency of experience with the information [84], personal beliefs [47], and socioeconomic status [117]. Intervention-centered factors include the phrasing of the intervention [17, 25] and prevalence of alternative information during intervention [50]. Though not a complete list, these findings demonstrate the complexity of evaluating the efficacy of misinformation tags. Figure 2 shows an illustration on how users who differ based on thinking style, political beliefs and attitudes, and personality traits could have different perceptions of misinformation tags, which can subsequently influence engagement with social media posts. To some users, the presence of a warning tag may have no impact on sharing behaviors while for others, the display of a warning tag may prevent engagement. While factors such as thinking styles and personality traits may show consistent effects on tag attitudes across misinformation topics, political beliefs and institutional trust may vary more often based on whether the topic is aligned with relevant group identities. In the present study, we hope to illuminate individual factors which may impact tag efficacy.

Given the wide range of individual factors which may impact a person's attitudes towards tags, it is unsurprising that prior research on warning label attitudes has resulted in mixed findings. As noted, Saltz et al. [100] found conservatives oppose labeling interventions while liberals supported labeling. Additionally, both conservatives and liberals attributed perceived labeling errors to bias in human judgments against their beliefs, rather than accidental algorithmic or human mistakes. Other works support the notion that who provides the warning label matters. Jia et al. [47] found warning labels provided by an algorithm, community, or third-party fact-checker were trusted by Democrats regardless of post ideology, while only algorithmic labels impacted Republican's

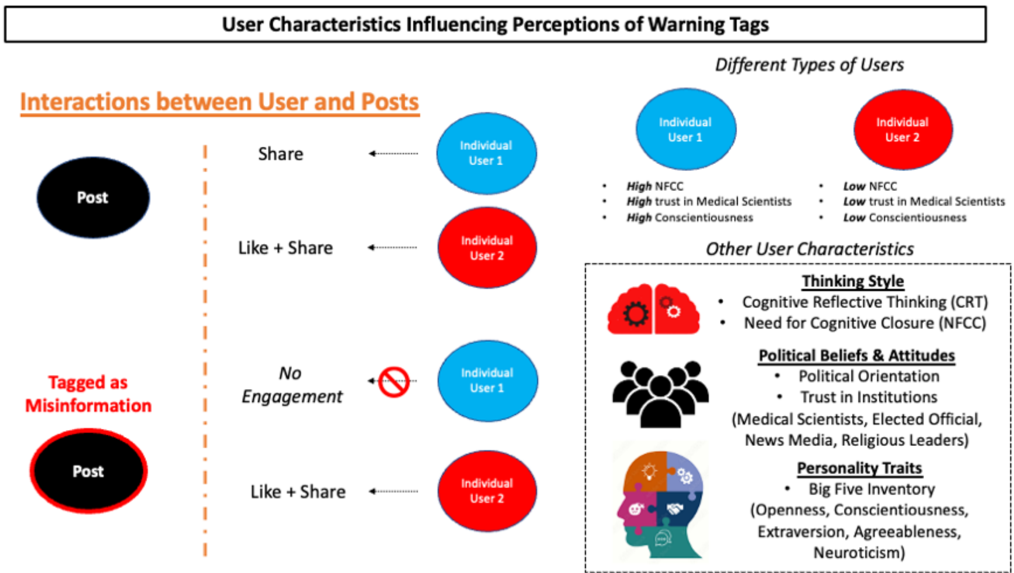

Fig. 2. Framework for how dispositional traits of users influence perceptions of misinformation warning tags and engagement behaviors.

belief of false conservative news. All intervention methods were effective for Republicans when the content was liberal-leaning.

Though it is clear that there are strong individual differences in the efficacy of tags, little work has been conducted on the personalization of warning tags. Bhuiyan et al. [10] investigated the impact of attention cue interventions (i.e., "nudges") on users exposed to general – not medical – true and false news headlines with consideration of political affiliation, media cynicism, and media skepticism characteristics. However, they do not examine personality traits such as those from BFI. The present study includes such characteristics, as well as provides a more nuanced attitude scale for evaluating how users may interact with warning tags. In a related study, Saltz et al. [101] identifies significant variance in individual preferences for warning tags. When asked about their attitudes towards warning tags, a plethora of concerns arose for users. Some users felt patronized by the social media platforms, thought the platforms were being too political, found the labels insufficient, or preferred that the posts be removed altogether. A recent attempt at personalizing moderation in the context of misinformation presents posts in a more culturally aligned way to the user's identifying political party Pretus et al. [91]. This study found that users are more receptive to posts that align with their political party, especially in polarizing contexts, providing "initial evidence that identity-based interventions may be more effective than identity-neutral alternatives for addressing partisan misinformation". The implication is that personalization may be an effective direction to increase the efficacy of mitigation efforts like tags. In order to get there, we need to know what individual differences matter.

Though prior work largely attributes discrepancies in warning tag efficacy to political affiliation, there is increasing evidence that other characteristics better account for an individual's misinformation susceptibility and sharing behaviors [10, 56, 66, 90, 109]. While this work finds that political orientation is associated with attitudes towards warning tags, factors such as CRT, NFCC, BFI Conscientiousness, Trust in Medical Scientists, and Trust in Religious Leaders were also influential.

Such results are important to design and deploy misinformation warning labels that can meet the needs of diverse groups.

3 Method

Participants from Amazon Mechanical Turk (MTurk) and SONA – an undergraduate population – filled out survey scales to determine their political affiliation, trust in institutions, BFI scores, CRT, and NFCC. The resulting scores were then analyzed with respect to the attitude scale found in Tables 5 and 7. Multivariate regression analysis was used to determine which individual characteristics best predict respondent attitudes towards warning tag labels.

3.1 Data Collection

The first set of respondents were recruited from SONA (n = 112), a pool of undergraduate students from a large university in the United States. The mean age of the SONA respondents was 21.6, of which 71.4% identified as female. For the MTurk sample (n = 132), the mean age was 37.3 years old and 42.4% identified as female. All respondents were anonymous. Respondents who failed an attention check and/or finished under the 10th percentile of completion time (<4.5 minutes, median completion time = 24 minutes) were excluded from analysis. Compensation varied between sources: MTurk workers were compensated financially at standard survey-taking rates, while SONA respondents were given course credits. Additional variables were included in the SONA sample to further examine effects from personality factors (i.e., openness, agreeableness, neuroticism, and extroversion). See Appendix Table 4 for detailed descriptions of the samples used in the present analysis.

3.2 Variable Descriptions

To measure attitudes and self-described behaviors towards misinformation warning tags – the main outcomes of this study – a tag attitude scale was adapted from existing scales used to assess attitudes towards explainable AI explanations (these provide justifications for AI-decisions, similar to algorithmically-generated tags) [105]. The tag scale consisted of 9 individual, 5-point likert scale items ranging from strongly disagree to strongly agree. Participants first saw the statement "When I see a social media post tagged as potential misinformation..." followed by each item. An image of a warning tag within a newsfeed was also shown as an example to respondents (See Figure 1). An aggregate (overall) score was calculated by averaging across items. The full scale can be viewed in Table 7. A higher scale score indicates more favorable attitudes towards misinformation tags. The first item "I usually ignore the tag" was reversed coded since it expresses negative sentiment towards tags. Tag attitude scale items showed acceptable consistency within the merged SONA and MTurk samples (Cronbach's alpha = 0.659) and was not normally distributed (Shapiro-Wilk $p < 0.01$). Within the SONA sample, tag attitude items were more consistent (Cronbach's alpha = 0.752) and normally distributed (Shapiro-Wilk = 0.176) but showed less consistency (Cronbach's alpha = 0.546) and not normally distributed (Shapiro-Wilk $p < 0.01$) for the MTurk sample.

To assess individual differences, the following scales were used:

- *Political orientation* – One question asking participants to identify their political beliefs, with 1 = Strong Democrat/Liberal to 6 = Strong Republican/Conservative.
- *Trust in Institutions* – One 4-item scale adapted from the 2019 Pew Research Center's American Trends Panel survey [28] asking respondents "How much confidence, if any, do you have in each of the following to act in the best interests of the public?". The institutions asked about were elected officials, news media, medical scientists, and religious leaders with response options ranging from 1 = No confidence at all to 4 = A great deal.

- *Big Five Inventory* – One 41-item scale from the Big Five Inventory (BFI) was used to evaluate participants across the personality dimension of Extroversion, Neuroticism, Agreeableness, Openness, and Conscientiousness [48, 49, 94]. Only the subscale for Conscientiousness was given to participants in the MTurk samples due to length constraints.
- *Cognitive Reflection Test* – The Cognitive Reflection Test (CRT) was employed to assess cognitive reflection using three questions. Each question initially suggests an "intuitive" answer, but arriving at the correct solution requires reflective thought. For instance, Question 1 states: "A bat and a ball together cost $1.10. The bat costs $1.00 more than the ball. How much does the ball cost?" The intuitive answer is 10 cents, but the correct answer is 5 cents. A higher number of correct answers indicates a higher CRT [27].
- *Need for Cognitive Closure* – A shortened 15-item scale [96], adapted from the original 42-item scale [119], was utilized to assess NFCC.

Bivariate correlations were run between all trait variables and each individual tag item. Tag items were not normally distributed, therefore the Kendall's Tau correlation coefficient was used. Correlations were also run between tested trait variables and the composite tag scale. Since the composite tag scale was normally distributed, Pearson's correlation was used with traits that were also normally distributed, while Kendall's Tau was used for traits with non-normal distributions. Multiple regression was also run between tested traits and the composite tag scale as the dependent variable. In order to account for potential multicollinearity of independent variables, a Shapley Value regression was run to show the variance contributed by each tested trait. A Shapley Value regression assesses the relative importance of predictors by computing all possible combinations of independent variables within a regression model and recording how much the R-sq changes with the addition or subtraction of each variable [12]. Shapley coefficients are standardized as they reflect the proportion of model variance (R-sq) attributed to each independent variable. See the following for further description and related applications of Shapley Value regressions: [20, 37, 44, 55, 56, 79, 107]. Regression model results were used to evaluate the strongest predictors of tag attitudes across thinking styles (CRT, NFCC), political beliefs and attitudes (political orientation, trust in institutions), and personality traits (BFI Extroversion, Neuroticism, Agreeableness, Openness, and Conscientiousness).

4 Results

First, we present bivariate correlations of tested traits (BFI, NFCC, etc.) and tag attitudes (aggregate score and individual items). Bivariate correlations were used to assess effects between the tested predictors and the tag attitudes independent of covariates. Since researchers or developers may not always have access to all available trait measures from users, we believe these results can reveal effective proxy measures. Further, the bivariate correlations provide further granularity to our results by showing how the predictor variables are associated with individual items from the tag attitude scale.

We then present multiple regression results, which include Shapley value coefficients to break down the top factors predicting warning tag attitudes. Regression modeling and Shapley coefficients were used to assess and rank the strength of tested predictors when controlled for each other. The strongest predictors are those that remain statistically significant when controlled for other factors in a multiple regression model and explain the largest proportion of the variance for tag attitudes based on Shapley coefficients. Effects from contextual differences of recruiting platforms (undergraduates from SONA pool vs. workers on Amazon Mechanical Turk) and respondent age were also tested as predictor variables. If recruiting platform shows a statistically significant effect when controlled for all tested user characteristics, this would suggest that contextual factors associated with the platforms further influence warning tag attitudes.

4.1 Predicting Tag Attitudes

Table 1 shows bivariate correlations for the merged MTurk and SONA samples. See appendix Tables 5 and 7 for bivariate correlations for each sample separately. NFCC is positively correlated with the aggregated tag attitude scale ($r = 0.267$, $p < 0.01$) and shows significant correlations with almost every tag item ($p < 0.01$). Trust in medical scientists is also positively correlated with the tag attitude scale ($r = 0.178$, $p < 0.01$) and significantly correlated with 6 of the 9 tag items ($p < 0.05$). Most items show a positive correlation with Trust in Medical Scientists, except for the item "I usually ignore the tag." Political orientation is the last variable to show a significant correlation with the aggregated scale ($r = -0.152$, $p < 0.01$). These results show that political conservatism was positively correlated with agreeing with the item "I usually ignore the tag" ($r = 0.179$, $p < 0.01$) while political liberalism was positively correlated with agreeing with items such as "I am less likely to read the post" ($r = -0.152$, $p < 0.01$) and "I approach the post content with suspicion or hesitancy" ($r = -0.157$, $p < 0.01$). Despite not showing significant correlations with the aggregated tag scale, all tested traits showed at least two statistically significant correlations with individual items. For the item "I usually ignore the tag," CRT and BFI Conscientiousness were negatively correlated ($r = -0.111$, $p < 0.05$; $r = -0.214$, $p < 0.01$) while trust in elected officials and religious leaders were positively correlated ($r = 0.154$, $p < 0.01$; $r = 0.203$, $p < 0.01$). Those with greater Trust in Elected Officials, News Media, and Medical Scientists were more likely to agree with the item "I evaluate the post content based on who the author is" ($r = 0.156$, $p < 0.01$; $r = 0.116$, $p < 0.05$; $r = 0.161$, $p < 0.01$). BFI Conscientiousness was negatively correlated with agreeing with the item "I trust that the post is, in fact, misinformation" ($r = -0.100$, $p < 0.05$) while age and Trust in Elected Officials, News Media, and Religious Leaders were positively correlated ($r = 0.124$, $p < 0.05$; $r = 0.175$, $p < 0.01$; $r = 0.136$, $p < 0.05$; $r = 0.110$, $p < 0.05$). When examining BFI traits only measured in the SONA sample (see Appendix Table 5), positive correlations were shown between the composite tag attitude scale and BFI traits for Openness ($r = 0.40$, $p < 0.01$), Conscientiousness ($r = 0.19$, $p < 0.05$), and Agreeableness ($r = 0.29$, $p < 0.01$).

Multiple regression model results with Shapley coefficients for both SONA and Mturk samples are shown in Table 2. Predictor variables for dispositional traits were selected based on being measured in both samples. The effect of recruitment platform is also tested ("Platform: SONA"), which shows the difference of tag attitude effects associated with the SONA respondents compared to MTurk. Respondent age is included as well to control for age-related effects. The current model is shown to be a solid predictor of attitude variance (R^2 of 0.23, adjusted R^2 of 0.20) and shows that NFCC is positively associated with favorable tag attitudes and is highly statistically significant ($\beta = 0.21$, $p < 0.001$) when controlled for all other variables. NFCC is also the strongest predictor in the model by explaining over a third of the total R^2 (36.3%). Additionally, Trust in Medical Scientists was positively associated with favorable tag attitudes ($\beta = 0.15$, $p < 0.001$) and is the second strongest predictor, contributing to 22.5% of variance. Both political conservatism ($\beta = -0.07$, $p < 0.01$) and Trust in Religious Leaders ($\beta = -0.10$, $p < 0.05$) were negatively associated with the aggregated attitude scale when controlled for tested predictors. However, political conservatism explained over twice the variance of tag attitudes compared to Trust in Religious Leaders (16.9% vs. 6.5%). Recruiting platform also showed a statistically significant effect, where respondents recruited on SONA were less likely to have favorable tag attitudes compared to MTurk workers ($\beta = -0.22$, $p < 0.05$). This effect explains up to 6.4% of tag attitude variance when controlled for all tested predictors.

Table 1. Bivariate correlations between tested variables and tag attitude items – SONA and MTurk combined (n=244)

	Overall	"I usually ignore the tag"	"I am less likely to read the post"	"I am less likely to engage with the post through sharing, replying, or liking"	"I evaluate the post content based on who the author is"	"I am more likely to research the post's topic in order to evaluate the content myself"	"I approach the post content with suspicion or hesitancy"	"I view the author of the post in a negative way"	"I trust that the post is, in fact, misinformation"	"I appreciate the social media platform's attempt to warn me about potential misinformation"
CRT	0.029	**-0.111***	-0.001	-0.022	0.060	0.026	0.102	-0.020	-0.081	0.005
NFCC	**0.267****	**0.250****	**0.186****	**0.171****	**0.156****	0.077	**0.186****	**0.296****	**0.169****	**0.163****
Conscientious	-0.013	**-0.214****	-0.053	-0.015	-0.002	-0.026	**-0.099***	**-0.118***	**-0.100***	0.067
Politics	**-0.152****	**0.179****	**-0.152****	-0.100	0.037	-0.018	**-0.157****	-0.081	**-0.112***	**-0.140****
Trust Elected	0.086	**0.154****	0.049	-0.046	**0.156****	0.103	-0.023	**0.185****	**0.175****	0.025
Trust News Media	0.074	0.066	0.061	-0.012	**0.116***	0.094	0.003	0.096	**0.136****	0.081
Trust Med Sci	**0.178****	**-0.139***	0.011	**0.199****	**0.161****	**0.180****	**0.221****	-0.042	0.035	**0.222****
Trust Religious	-0.028	**0.203****	-0.089	-0.099	0.102	0.090	-0.052	-0.001	**0.110***	-0.013
Age	0.033	0.029	-0.036	-0.082	0.089	0.072	-0.068	**0.176****	**0.124****	-0.034

Attitude responses prefaced with "When I see a social media post tagged as potential misinformation..."
Significance codes: $^{*}p < 0.05$, $^{**}p < 0.01$

Table 2. Multiple regression – Tag attitude aggregate as dependent variable (MTurk & SONA)

	β coef	Shapley R^2	std err	t-value	p-value
Politics	**-0.07**	**16.9%**	**0.02**	**-3.32**	**<0.01**
CRT	0.01	0.5%	0.03	0.53	0.59
NFCC	**0.21**	**36.3%**	**0.04**	**5.04**	**<0.001**
Conscientiousness	-0.02	0.2%	0.06	-0.29	0.77
Trust Elected	0.06	4.4%	0.05	1.20	0.23
Trust News Media	0.04	4.9%	0.04	0.99	0.33
Trust Med Sci	**0.15**	**22.5%**	**0.04**	**-2.45**	**0.02**
Trust Religious	**-0.10**	**6.5%**	**0.04**	**-2.45**	**0.02**
Age	0.00	1.3%	0.00	-1.23	0.22
Platform: SONA	**-0.22**	**6.4%**	**0.10**	**-2.06**	**0.04**
R^2	0.23				
Adj R^2	0.20				

4.2 SONA and MTurk Comparison

Both SONA and MTurk (Appendix 5 and 7) showed correlations between Conscientiousness and the aggregate tag attitude score, however, the effects were in different directions. For SONA, higher conscientiousness was correlated with more favorable tag attitudes while among MTurk workers conscientiousness was negatively correlated. The samples did not share any other correlations with respect to the aggregate tag attitude – though they did with certain individual attitude measures. While NFCC showed no significant overall effects among SONA participants, among MTurkers NFCC was positively correlated with the aggregate tag attitude scale ($r = 0.447$, $p < 0.01$) as well as all individual items. Significant effects found in the SONA sample but not in the MTurk sample include Politics, Trust in Medical Scientists, and Trust in Religious Leaders. BFI Openness, Agreeableness, Extraversion, and Neuroticism were not tested in the MTurk sample.

Comparing multiple regression results (Appendix 6 and 8), both found Politics and Trust in Medical Scientists to be significant predictors of tag attitudes when controlling for the other tested variables. The SONA sample also found Openness to be significant, which was not available in the MTurk sample. Further, the MTurk sample found NFCC, Conscientiousness, and Trust in Religious Leaders to be statistically significant while the SONA sample did not. It is possible that the SONA regression model may show fewer significant effects than MTurk due to having a greater number of examined covariates.

4.3 Summary of Results

In summary, the multiple regression model and Shapley coefficients show that NFCC was the strongest predictor of tag attitudes when controlled for all tested variables, followed by Trust in Medical Scientists and political orientation. Trust in Religious Leaders and recruiting platform also showed significant effects on tag attitudes, although the effects were weaker. Bivariate correlations show that every tested factor had a statistically significant effect with individual tag attitude items, which has implications for future warning tag design. When examining the SONA sample only, the regression shows that Politics, Openness, and Trust in Medical Scientists were significant ($p < 0.05$) when controlling for other variables. Within the MTurk sample, NFCC was the strongest predictor by a large margin when controlled for all other variables, followed by Politics, Trust in Medical Scientists, and Trust in Religious Leaders. Differences between the two distinct samples tested in this study further emphasize the importance of individual differences in understanding the impact of misinformation mitigation strategies. Hypotheses and support for these hypotheses are summarized in Table 3.

Table 3. Tested variables and hypotheses for attitudes on misinformation tags.

Traits	**Hypotheses**	**Present Findings***
Political Beliefs and Attitudes	H1: Political conservatism is negatively associated with favorable tag attitudes H2: Trust in Medical Scientists, Elected Officials, and News Media will be positively associated with favorable tag attitudes H3: Trust in Religious Leaders is negatively associated with favorable tag attitudes	H1: Full Support H2: Partial Support – Only Trust in Medical Scientists showed expected effects H3: Full Support
Thinking Styles	H4: CRT is positively associated with favorable tag attitudes H5: NFCC is positively associated with favorable tag attitudes	H4: Not Supported H5: Full Support
Personality	H6: BFI Conscientiousness will be positively associated with favorable tag attitudes H7: BFI Openness and Neuroticism will be positively associated with favorable tag attitudes H8: BFI Agreeableness and Extraversion will be negatively associated with favorable tag attitudes	H6: Partial Support – Only showed significant effect in MTurk model and was negative H7: Partial Support – Only BFI Openness shows expected effect H8: Not Supported

*Note: Support for hypotheses are based on results from the multiple regression models.

5 Discussion

The present study supports the premise that cognitive factors, beliefs, and personality traits impact how a person views and interacts with misinformation warning tags. Specifically, our results indicate that several individual factors – including Politics, NFCC, Openness, Conscientiousness, Agreeableness, Trust in Medical Scientists, and Trust in Religious Leaders – predict a person's attitudes towards misinformation warning tags. When controlled for effects from covariates and multicollinearity, NFCC, political orientation, and Trust in Medical Scientists were the strongest predictors for tag attitude variance among our merged sample. These results may explain discrepancies in past literature on the effectiveness of tags as a mitigation strategy, as certain tags may work for some people but not others. Our results highlight the premise that designing tailored misinformation interventions as opposed to taking one-size-fits-all strategies may benefit designers seeking to curb misinformation for diverse groups.

Predictably, we find differences in characteristics between MTurk and SONA samples. Differences seen between the samples strengthen the case that different populations may have different needs for warning tags, and thus may require personalization. The SONA population presents a younger, female-majority cohort of undergraduate students motivated by course extra credit, while MTurk presents a more moderate, older and balanced gender population motivated by financial gain. Improved predictions of a user's attitude towards warning labels can be leveraged by future designers as a means to personalize interventions that work best for specific user groups. In the remainder of this paper, we will discuss the implications of our results and provide recommendations for designing more effective misinformation warning tags.

5.1 Individual Factors Predict Warning Tag Attitudes

Political orientation is one indicator of warning tag attitudes, but does not show the whole picture. We echo prior results showing political orientation plays a large role in attitudes towards warning tags, confirming H1 [8, 100]. Indeed, Politics was the strongest predictor of tag attitudes for the SONA dataset. For MTurk, however, Politics was a weaker predictor of tag attitudes compared to other significant effects. In all samples, liberal-leaning respondents were more likely to have a positive attitude towards warning tags, while conservative-leaning respondents were likely to have negative attitudes.

Importantly, while Politics may be an effective indicator of tag attitudes in some cases, our findings indicate that other factors such as Openness and NFCC are also strong predictors of a respondent's attitude towards warning tags. The inclusion of these other factors can provide a more comprehensive understanding of how different groups may think about and behave toward warning tags. In a more general sense, these results imply that the research community should look beyond well-studied factors like political orientation and examine other personal characteristics – including a person's cognitive and psychological profile and trust judgments – to deepen our understanding of how they may interact with misinformation mitigation strategies like tags.

The sources of information a person trusts may be an important predictor of their attitude towards tags. Those who had greater trust in medical scientists were more likely to have positive attitudes about warning tags for both the merged and SONA-only samples. This confirms part of H2: when examining individual tag attitudes, respondents who had higher trust in medical scientists report being more likely to perceive a post as misinformation and less likely to engage with a post when tagged as misinformation. As prior work shows that those who use factual news sources have higher trust in medical scientists [65], the positive associations with tag attitudes may reflect a higher respect for institutional guidance and public health recommendations. More generally, it may also reflect a person's preference for expert judgment and evidence-based decision-making. Trust in the other tested institutions, News Media and Elected Officials, showed no significant effects on the aggregated tag attitude score.

Trust in Religious Leaders was negatively associated with favorable tag attitudes among the merged samples, supporting H3. One explanation for this finding may be due to differences in the criteria by which credibility of information is assessed: people who have higher trust in religious leaders may prioritize spiritual authorities over scientific authorities and distrust secular institutions like social media platforms. Though prior work has found that religious beliefs do not indicate conspiracy theory belief [46], differences in credibility assessment and trust in institutions has important implications on misinformation intervention design.

Information assessment traits may be strong predictors of tag attitudes. CRT did not show a significant effect on the aggregate attitudes score when controlled for our tested covariates, leading us to reject H4. When examining bivariate correlations, CRT was only shown to have a negative effect with the individual item "I usually ignore the tag". While prior work shows that respondents higher in CRT tend to be more capable of distinguishing real from fake news potentially due to their propensity to research the topics themselves [82] [8], our findings suggest that CRT is not a major predictor for general attitudes towards tag warning systems. The discrepancy between our results and previous work may be further explained by differences in outcome measures, as CRT may be a relevant predictor for evaluating the veracity of news headlines but less effective when examining other measures of propagation behaviors on social media [37].

We also found that respondents who have a desire for certainty and predictability (i.e. high NFCC) are more likely to have positive attitudes towards warning tags, supporting H5. Intuitively, this makes sense: warning tags have the potential to bring clarity on whether a post's information can be relied upon. Interestingly, we did not find significant effects between NFCC and tag attitudes for the SONA sample. This distinction between the SONA and MTurk sample warrants further investigation, though we suspect it may be due to the overall higher tolerance for uncertainty observed by the university students as compared to our more general MTurk population (see Appendix Table 4 for mean comparisons).

Personality matters: BFI Conscientiousness was negatively associated with tag attitudes among Mturk workers, and Openness was positively associated for SONA participants when controlling for tested covariates. These findings partially support H6, as Conscientiousness showed a significant association with tag attitudes in the MTurk-only regression, but was in the opposite direction than expected. This effect could stem from a desire for self-sufficiency or a preference to avoid outside judgments if they believe they can handle the task themselves and believe those providing interventions may have goals counter to their own. If the less favorable views of tags among Conscientious individuals is driven by stronger confidence in information assessment abilities, then the present results may still be consistent with previous work showing associations between Conscientiousness with better news discernment and lower tendency to share misinformation [16, 66].

We did find mixed results, however, when comparing samples. Conscientiousness was negatively correlated with aggregate tag attitudes in the MTurk sample, but showed a positive correlation for SONA respondents. It is also worth noting that the effect from Conscientiousness from MTurkers remained significant when controlled for the influence from the other tested variables while it no longer showed an effect in the SONA regression model. This may indicate that for MTurk workers, Conscientiousness is a more influential trait for misinformation-related outcome measures. This is further suggested in related work, which showed that Conscientiousness was the strongest predictor for misinformation detection accuracy among MTurkers [56]. The difference in the direction of correlation effects may also be attributed to contextual differences when recruiting respondents from MTurk compared to SONA. For MTurk workers, Conscientiousness may be a more influential factor because being orderly and cautious would be more useful attributes when completing other tasks on the platform, which typically involve an attention to detail to receive approval for the work (e.g., correctly labeling images to be used as training data). In contrast, undergraduates enrolled in a university course may place less importance on being Conscientious, as they are not receiving monetary compensation for completing tasks requiring attention to detail. Undergraduates may also have more favorable views of social media companies in general, which could result in more favorable views of tagging interventions. Results in Table 5 support this possibility, where Conscientious was positively correlated with the statement "I appreciate the social media platform's attempt to warn me about potential misinformation" among SONA respondents.

Our findings partially support H7, as Openness is positively associated with favorable tag attitudes in the SONA-only regression. However, there are no effects observed with Neuroticism. This is consistent with previous work showing that Openness is associated with better news discernment and skepticism towards myths [8, 16, 111]. We believe openness in this case may be reflective of how willing a person is to accept an outside judgment of the credibility of information. It is important to keep in mind that Openness was not tested in the MTurk sample, and thus comparisons cannot be drawn between samples. Our findings are inconsistent with [8], as we do not observe a significant effect between Neuroticism and tag attitudes. Similar to the effects previously described for CRT, differences in outcome measures may explain discrepancies between the current findings and prior work. In other words, Neuroticism may be an effective predictor for evaluating truthfulness of news headlines but may not be a reliable factor that influences general attitudes towards tag warning systems.

Lastly, there were no significant effects observed in the regression analysis for both Agreeableness and Extraversion on tag attitudes. Therefore, we reject H8. While we did not observe significant effects from Agreeableness when controlled for our tested covariates, our results show significant effects with the aggregate attitude score and individual items when using bivariate correlations. These findings suggest that respondents who are more Agreeable – i.e. they have a greater care for social harmony and cohesion – may have more positive attitudes towards some aspects of warning label interventions. This intuitively makes sense as well: misinformation warning tags are, in a general sense, pro-social, and may be seen as a way to promote truth, show care for one's community, and reduce the negative impacts of spreading false information. This aligns with a previous work claiming users high in agreeableness are more likely to validate news prior to sharing said news [102] but diverges with work finding no correlation between agreeableness and sharing flagged content [8]. Since Agreeableness did not show significant effects in the multiple regression model, this implies that it is less important to consider than other characteristics.

5.2 Design Implications

In this study, we found that certain traits can predict people's attitudes and self-described behaviors towards misinformation warning tag interventions. These results have implications for the design of misinformation mitigation solutions like warning tags. Most importantly, our results highlight the need for personalized and culturally-relevant warning tag interventions, as one-size-fits-all approaches to prevent misinformation uptake and spread may fail if they don't harmonize with the specific needs and motivations of a particular individual. These include structuring messages for audiences with different psychological, cognitive, and information-assessment traits as well as using an understanding of credibility signaling to help people who trust different authority sources heed similar warnings. Concretely, this could be realized in the real world by offering curated interaction experiences on social media sites like Facebook or X to people with different trait profiles, such as showing certain information for some groups but not for others. For developers interested in tagging interventions, labels can be further tailored to trait profiles by adapting text wording, graphics used (e.g., red exclamation mark), and whether to cite authorities or official institutions (e.g., World Health Organization). Taken together, our findings provide actionable direction for future warning tag design and misinformation mitigation solution design more generally.

Designing For Cognitive and Informational Assessment Traits. Our results imply that cognitive and information assessment traits like Need for Cognitive Closure (NFCC) may meaningfully impact a person's attitude and behavior towards misinformation warning tags. The general implication is that these traits should be accounted for when designing warning tags and other misinformation mitigation solutions. The results of our study can provide some direction for designers, however, further research is needed on how exactly to design for people with high and low NFCC. As we may not always be able to detect a person's cognitive and informational assessment traits, it may be best to design for the anti-tag attitude cases.

People high in NFCC prefer clear, unambiguous information, and thus we saw in our study that these individuals (particularly in our MTurk sample) had more favorable attitudes towards warning tags, and thus warning tags may be appropriate and effective for this population. A more challenging design problem is designing for people *low* in NFCC. For these people, messaging regarding the importance of certainty and clarity on the truthfulness of information and the implications of spreading false information may be helpful. In this way, even people who tend to be less concerned about ambiguity may engage in more information discernment behaviors.

Our results suggest that people high in CRT are less likely to ignore the warning (MTurk) and are more likely to research the topic themselves (SONA). Warning tag interventions may capitalize on the user's interest in self determination by presenting relevant research articles or news from reputable sources. For misinformation in the health domain specifically, harder moderation approaches with detailed explanation on why the determination was made may be appropriate for these users. Users lower in CRT may warrant behavioral intervention or harder ("hands-off") moderation approaches. Since these users tend to ignore warning tags and are less likely to conduct their own research, prompts requesting behavior (for example, *"Are you sure you want to share this potentially false post?"*) may be helpful. Severe cases of misinformation may need to be completely removed for these users.

Designing For Personality. Our results indicate that tailoring misinformation warning tags and other mitigation solutions based on a person's personality may be an effective approach. Prior work has shown connections between personality and misinformation vulnerability, and our study shows consistent results [4, 43, 56, 63, 66]. Recent work in social media research has shown that personality can be accurately detected from a person's social media behavior [26] – we suggest that similar approaches may be used as a means to inform the tailoring of personality-based warning tags.

People high in Openness may benefit from warning tags as a means to show them external perspectives. To design for people low in Openness, messaging emphasizing the importance of exploring new, diverse, or more objective perspectives to their own may be necessary for warning tag acceptance. Considering people low in Openness are less interested in conducting their own research, providing justification for misinformation as part of the label may provide a prudent safe guard for these more susceptible users. Other priming interventions, such as "accuracy prompts" which asks users to rate a neutral headline to encourage users to scrutinize future articles [87], may also be pertinent.

Similar to Openness, people high in Agreeableness may benefit from warning tags inherently. For people high in Agreeableness, emphasis on the pro-social nature of warnings may increase their effectiveness. For people low in Agreeableness, prompts to increase empathy, orient the person to understand the impact of misinformation on vulnerable communities, and emphasize compassion may be helpful to increase warning tag efficacy.

People high in Conscientiousness may benefit from warning tags emphasizing the importance of sharing verified information, whereas those low in Conscientiousness may be less organized, so simple messaging and clear visuals that require less detail orientation may be helpful. Those high in Conscientiousness and more trusting of social media platforms (such as from our undergraduate sample) will likely benefit from warning tag interventions as they are. Users similar to those in the MTurk population that are also high in Conscientiousness trust warning tags less and may require interventions that increase trust for these particular users. Providing explanations about the underlying decision making process (algorithmic, community, 3rd party, etc.), could increase user trust with platforms as well. We posit 3rd party moderators may be particularly effective at increasing platform trust with these users.

Designing For Trust And Credibility Based On Identity And Belief. Our study results suggest that the informational sources one trusts as well as their belief systems impact a person's attitudes towards misinformation warning tags.

In the case of political orientation, self-identified liberals were generally more positive about warning tags than conservatives. Understanding *why* there may be differences based on political ideology may reflect differences in media consumption habits, including trust in mainstream media and the perceived bias of misinformation-identifying algorithms, as well as the bias of the misinformation itself (i.e. if the misinformation supports republican or democratic-led claims). Designing warning tags based on partisan divides is a challenge when one wants to remain objective in the discernment of trust based on demonstrable fact, however, some general principles may apply. There is evidence to suggest that republicans and democrats may differ in their value judgments [34], and thus tailored tags that appeal to each group's value systems may be a good place to start. For example, liberals may respond better to appeals to institutional authority, collective well-being, or scientific argument, whereas conservatives may respond better to appeals to individual freedom and personal responsibility. In both cases, endorsement by authority figures from their respective parties may help increase the effectiveness of warning tags.

In a similar vein, we find that the specific institutions a person trusts impacts their warning tag attitudes. Our study found people high in Trust in Medical Scientists had more favorable tag attitudes, while people with high Trust in Religious Leaders had less favorable tag attitudes. These results imply the importance of mentioning authority figures that are credible to users as well as using messaging that aligns with a person's reasoning priorities. For example, warning tags for people high in Trust in Medical Scientists may benefit from logical arguments based in scientific reasoning, while people low in Trust in Medical Scientists but high in trust in a different institution – such as Religious Leaders – may benefit from appeals to specific spiritual values, culturally-specific messaging, and moral responsibility. Previous work showing that medical authority is often used to propagate misinformation (e.g., stating a false health treatment is endorsed by a medical professional) [36, 39] should also be taken into consideration when designing tag interventions, as this increases the difficulty for algorithms to detect misinformation. In these scenarios, tag labels that hedge or use less certain language (e.g., this may contain misinformation) could be more effective by mitigating effects from potentially labeling accurate scientific information as misinformation. For people high in Trust in Religious Leaders specifically, however, it is possible that warning tags may not be the best intervention. These users may be more likely to ignore tags and may even have *more* positive views of authors of potential misinformation. More generally, interventions may also attempt to improve trust by catering to a person's preference of moderation group (community, third party, algorithmic) to best determine what should be removed, as long as it remains objective.

Ethical Implications of Tailored Warning Tags. Our work has significant ethical implications. Prior research demonstrates that social media behavior, such as Facebook likes, can reveal psychological characteristics, including Big Five personality traits [61]. Tag warning systems can leverage this insight by classifying users and tailoring interventions based on their online behaviors. While this approach presents a promising opportunity to personalize misinformation warnings at scale, it raises important ethical concerns, particularly regarding user consent and agency. Personalized interventions could be implemented without the explicit consent of users, relying on assumptions about their preferences and vulnerabilities. Although these interventions aim to assist users in identifying misinformation, it is crucial to preserve user autonomy and privacy. Users should have the ability to opt out of personalized interventions or biasing algorithms, ensuring they retain control over their online experience. Future work should explore the default settings for platforms employing such interventions and consider the implications of involving users in decisions about their desired social media experience. Furthermore, it is essential to address the risk of misusing personalization data, especially for vulnerable populations. Careful safeguards must be implemented to ensure equitable access to tagging information and to mitigate algorithmic biases that could disproportionately affect marginalized groups. Lastly, it is worth noting that those who might benefit most from these interventions – such as individuals with a high distrust of institutions – may also be the least likely to adopt them. Addressing this challenge remains an open question. We argue that, above all, users must have the agency to decide what level and type of personalization is appropriate for them, striking a balance between effective misinformation mitigation and respect for individual autonomy.

Broader Implications. Up until this point, we have primarily focused on individual differences with regards to warning tag interventions for online misinformation specifically. We expect that many of the implications of this study – particularly the need for personalization and culturally-relevant design – to generalize to other types of misinformation interventions or even to other domains where algorithmically-determined decisions need to be trusted or understood by a particular user, such as explainable AI. In particular, we expect that many of the design implications detailed above, including how to design for cognitive and informational traits, personality, and beliefs about trust and credibility, may generalize to other contexts. The concept of personalization is not new – personalization and tailoring has been commonplace in public health and advertising for decades [62], and has become a major topic in human-computer (particularly human-AI) interaction [41, 57, 81, 103, 115]. Our results support this direction of work and may be used by practitioners in a variety of HCI domains as a means to better understand and design for diverse populations.

6 Limitations and Future Work

This study was not without limitations. This study was conducted in the United States, as such the results may not generalize to other populations. The SONA results may be more applicable to younger/liberal leaning populations, while MTurk results may can be applied to more general populations. MTurk workers were motivated via financial incentives while SONA participants were motivated by class credits, resulting in potentially different values when taking the survey, and though unlikely, it is possible compensation may have biased results. When assessing the validity of our tags attitude scale, our results show a lower Cronbach's alpha value (0.546) for the MTurk sample. While the individual items in our scale still reflect specific types of reactions to warning tags (e.g., ignore it) that can inform interventions, further refinement may be warranted to measure general attitudes towards tag warning systems more consistently.

There is always the risk of dishonest answers from respondents who partook in this study, though the study was voluntary and anonymous to limit this risk. Since data is self-reported, we do not know how actual behavior in real world settings may differ from the reported attitudes in this study. Mosleh at al. [74] find self-reported willingness to share data translates to actual sharing behavior in the context of sharing news. We make the assumption that such indications hold for other interaction behavior in our survey but we encourage future work to investigate if self-reported actions such as ignoring tags, likelihood to read a post, post author evaluation in the presence of a warning tag, suspicion or hesitancy in the presence of a warning tag, and attitudes towards a platform in the presence of warning tags are consistent with these self-reported responses. In addition, self-reported political affiliation may have variation stemming from misperception of what it means to identify with a party [121]. This may in part explain some of the variance seen in the earlier results.

These limitations motivate potential future research and design studies. A user study testing the effectiveness of tag interventions "in the wild" to better assess real world behavior towards different types of tailored warning tags would be a natural next step. Future work should also consider adapting simulated social media newsfeeds [45] to further assess the effects from interventions and user characteristics on behavioral outcomes. In fact, the use of simulated newsfeeds and survey scales can reveal greater nuance in user reactions to tag warning systems by accounting for both behaviors and attitudes. Our results show differences between MTurk and SONA recruiting platforms when controlled for age and political orientation, implying that contextual differences (e.g., incentive) and other characteristics not analyzed in the current study (e.g., education level, income) may also influence tag attitudes. Further examination is needed to further explain attitude differences between recruiting platforms. As our results are based only on US samples, future research should consider sampling from other countries to assess cultural differences in warning tag attitudes; warning tags may then be personalized accordingly to these additional user populations. Beyond cultural differences, future work may investigate additional populations or characteristics to better understand how attitudes may differ.

7 Conclusion

In this study, we investigated how cognitive factors, beliefs, and personality impact attitudes towards misinformation warning tags. This is an important area of study, as online misinformation is rampant and warning tags may provide an effective mitigation solution. Using two distinct sample groups, we find that NFCC, Trust in Medical Scientists, and political liberalism are the strongest predictors of favorable attitudes towards tag warning systems while Trust in Religious Leaders is negative associated. When examining effects within each sample, we find evidence that Openness, Agreeableness, and CRT can also influence attitudes towards warning tags. These characteristics can be used to predict user attitudes towards warning labels and be leveraged to personalize misinformation warning tags, which can potentially increase their efficacy for diverse populations. Results are synthesized to provide direction for designing online misinformation interventions based on a person's cognitive traits, personality, and beliefs about trust and credibility. The implications of this study may generalize to other domains where algorithmically-determined decisions are deployed at-scale, including other human-computer interaction domains.

Acknowledgments

The authors would like to acknowledge Steven Dow, Chloe Lee, and Joseline Chang for supporting the data collection process. A special thanks to the Discovery Way Foundation for their generous usage of lab space.

References

[1] Jon D. Agley and Yunyu Xiao. 2020. Misinformation about COVID-19: evidence for differential latent profiles and a strong association with trust in science. *BMC Public Health* 21 (2020).

[2] Saifuddin Ahmed and Han Wei Tan. 2022. Personality and perspicacity: Role of personality traits and cognitive ability in political misinformation discernment and sharing behavior. *Personality and Individual Differences* (2022).

[3] Bushra Alsaadi and Dimah Alahmadi. 2021. The use of Persona towards human-centered design in health field: review of types and technologies. In *2021 International Conference on e-Health and Bioengineering (EHB)*. IEEE, 1–4.

[4] Irum Alvi and Niraja Saraswat. 2020. Information Processing – Heuristic vs Systematic and Susceptibility of Sharing COVID-19-related Fake News on Social Media. *Journal of Content, Community and Communication* 12 (2020).

[5] Adam Badawy, Emilio Ferrara, and Kristina Lerman. 2018. Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 258–265. doi:10.1109/ASONAM.2018.8508646

[6] João Pedro Baptista and Anabela Gradim. 2022. A working definition of fake news. *Encyclopedia* 2, 1 (2022).

[7] Pablo Barbera. 2018. Explaining the spread of misinformation on social media: evidence from the 2016 US presidential election. In *Symposium: Fake News and the PoliticsofMisinformation. APSA*.

[8] Dipto Barman, Kevin Koidl, and Owen Conlan. 2024. Discerning Individual Preferences for Identifying and Flagging Misinformation on Social Media. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*. 110–119.

[9] Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. 2015. Science vs conspiracy: Collective narratives in the age of misinformation. *PloS one* 10, 2 (2015), e0118093.

[10] Md Momen Bhuiyan, Michael Horning, Sang Won Lee, and Tanushree Mitra. 2021. Nudgecred: Supporting news credibility assessment on social media through nudges. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–30.

[11] Aengus Bridgman, Eric Merkley, Peter John Loewen, Taylor Owen, Derek Ruths, Lisa Teichmann, and Oleg Zhilin. 2020. The causes and consequences of COVID-19 misperceptions: Understanding the role of news and social media. *Harvard Kennedy School Misinformation Review* 1, 3 (2020).

[12] David V Budescu. 1993. Dominance analysis: a new approach to the problem of relative importance of predictors in multiple regression. *Psychological bulletin* 114, 3 (1993), 542.

[13] Timothy Callaghan, Ali Moghtaderi, Jennifer A Lueck, Peter J Hotez, Ulrich Strych, Avi Dor, Erika Franklin Fowler, and Matt Motta. 2020. Correlates and disparities of COVID-19 vaccine hesitancy. *Available at SSRN 3667971* (2020).

[14] Dustin P Calvillo, Ryan JB Garcia, Kiana Bertrand, and Tommi A Mayers. 2021. Personality factors and self-reported political news consumption predict susceptibility to political fake news. *Personality and individual differences* 174 (2021), 110666.

[15] Dustin P Calvillo, Alex León, and Abraham M Rutchick. 2024. Personality and misinformation. *Current opinion in psychology* 55 (2024), 101752.

[16] Dustin P Calvillo, Bryan J Ross, Ryan JB Garcia, Thomas J Smelter, and Abraham M Rutchick. 2020. Political ideology predicts perceptions of the threat of COVID-19 (and susceptibility to fake news about it). *Social Psychological and Personality Science* 11, 8 (2020), 1119–1128.

[17] Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. 2020. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political behavior* 42 (2020), 1073–1095.

[18] X Corp. 2021. *Updates to our work on COVID-19 vaccine misinformation.* https://blog.x.com/en_us/topics/company/2021/updates-to-our-work-on-covid-19-vaccine-misinformation

[19] X Corp. 2024. *Community Notes: a collaborative way to add helpful context to posts and keep people better informed.* https://communitynotes.x.com/guide/en/about/introduction

[20] Ian Covert and Su-In Lee. 2021. Improving kernelshap: Practical shapley value estimation using linear regression. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 3457–3465.

[21] Jonas De Keersmaecker, David Dunning, Gordon Pennycook, David G Rand, Carmen Sanchez, Christian Unkelbach, and Arne Roets. 2020. Investigating the robustness of the illusory truth effect across individual differences in cognitive

ability, need for cognitive closure, and cognitive style. *Personality and Social Psychology Bulletin* 46, 2 (2020), 204–215.

[22] Paula de Pando Mena. 2020. Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook. *Policy & Internet* (2020).

[23] James N. Druckman and Mary C. McGrath. 2019. The evidence for motivated reasoning in climate change preference formation. *Nature Climate Change* 9 (2019), 111–119.

[24] Alice H Eagly and Shelly Chaiken. 2007. The advantages of an inclusive definition of attitude. *Social cognition* 25, 5 (2007), 582–602.

[25] Ullrich K. H. Ecker, Stephan Lewandowsky, and David T W Tang. 2010. Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition* 38 (2010), 1087–1100.

[26] Golnoosh Farnadi, Geetha Sitaraman, Shanu Sushmita, Fabio Celli, Michal Kosinski, David Stillwell, Sergio Davalos, Marie-Francine Moens, and Martine De Cock. 2016. Computational personality recognition in social media. *User modeling and user-adapted interaction* 26 (2016), 109–142.

[27] Shane Frederick. 2005. Cognitive reflection and decision making. *Journal of Economic perspectives* 19, 4 (2005), 25–42.

[28] Cary Funk, Meg Hefferon, Brian Kennedy, and Courtney Johnson. 2019. Trust and mistrust in Americans' views of scientific experts. *Pew Research Center* 2 (2019), 1–96.

[29] Tarleton Gillespie. 2020. Content moderation, AI, and the question of scale. *Big Data & Society* 7, 2 (2020), 2053951720943234.

[30] Jeffrey Gottfried and Elisa Shearer. 2016. News use across social media platforms 2016. (2016).

[31] Rebecca Hofstein Grady, Peter H. Ditto, and Elizabeth F. Loftus. 2021. Nevertheless, partisanship persisted: fake news warnings help briefly, but bias returns with time. *Cognitive Research: Principles and Implications* 6 (2021).

[32] Andrew Guess, Brendan Nyhan, and Jason Reifler. 2018. Selective exposure to misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign. (2018).

[33] Andrew Markus Guess, Jonathan Nagler, and Joshua A. Tucker. 2019. Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances* 5 (2019).

[34] Jonathan Haidt. 2012. *The righteous mind: Why good people are divided by politics and religion.* Vintage.

[35] Oliver L. Haimson, Dan Delmonaco, Peipei Nie, and Andrea Wegner. 2021. Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas. *Proceedings of the ACM on Human-Computer Interaction* 5 (2021), 1 – 35.

[36] Michael Robert Haupt, Michelle Chiu, Joseline Chang, Zoe Li, Raphael Cuomo, and Tim K Mackey. 2023. Detecting nuance in conspiracy discourse: Advancing methods in infodemiology and communication science with machine learning and qualitative content coding. *Plos one* 18, 12 (2023), e0295414.

[37] Michael Robert Haupt, Raphael Cuomo, Tim K Mackey, and Seana Coulson. 2024. The role of narcissism and motivated reasoning on misinformation propagation. *Frontiers in Communication* 9 (2024), 1472631.

[38] Michael Robert Haupt, Alex Jinich-Diamant, Jiawei Li, Matthew Nali, and Tim K Mackey. 2021. Characterizing twitter user topics and communication network dynamics of the "Liberate" movement during COVID-19 using unsupervised machine learning and social network analysis. *Online Social Networks and Media* 21 (2021), 100114.

[39] Michael Robert Haupt, Jiawei Li, and Tim K Mackey. 2021. Identifying and characterizing scientific authority-related misinformation discourse about hydroxychloroquine on twitter using unsupervised machine learning. *Big Data & Society* 8, 1 (2021), 20539517211013843.

[40] Michael Robert Haupt, Staci Meredith Weiss, Michelle Chiu, Raphael Cuomo, Jason M Chein, and Tim Mackey. 2022. Psychological and situational profiles of social distance compliance during COVID-19. *Journal of Communication in Healthcare* 15, 1 (2022), 44–53.

[41] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9, 4 (2019), e1312.

[42] Jina Huh, Bum Chul Kwon, Sung-Hee Kim, Sukwon Lee, Jaegul Choo, Jihoon Kim, Min-Je Choi, and Ji Soo Yi. 2016. Personas in online health communities. *Journal of biomedical informatics* 63 (2016), 212–225.

[43] V. Indu and Sabu M. Thampi. 2021. A Systematic Review on the Influence of User Personality in Rumor and Misinformation Propagation Through Social Networks.

[44] Osnat Israeli. 2007. A Shapley-based decomposition of the R-square of a linear regression. *The Journal of Economic Inequality* 5 (2007), 199–212.

[45] Arvin Jagayat and Becky L Choma. 2024. A primer on open-source, experimental social media simulation software: Opportunities for misinformation research and beyond. *Current Opinion in Psychology* 55 (2024), 101726.

[46] Inga Jasinskaja-Lahti and Jolanda Jetten. 2019. Unpacking the relationship between religiosity and conspiracy beliefs in Australia. *The British journal of social psychology* (2019).

[47] 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 (2022), 1–27.
[48] Oliver P John, EM Donahue, and R_L Kentle. 1991. The big five inventory: Versions 4a and 54 [Technical Report]. *Berkeley: University of California, Institute of Personality and Social Research* (1991).
[49] Oliver P John, Laura P Naumann, and Christopher J Soto. 2008. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research* 3, 2 (2008), 114–158.
[50] Irene P. Kan, Kendra L. Pizzonia, Anna Drummey, and Eli J. V. Mikkelsen. 2021. Exploring factors that mitigate the continued influence of misinformation. *Cognitive Research: Principles and Implications* 6 (2021).
[51] Robert Kaufman, Aaron Broukhim, David Kirsh, and Nadir Weibel. 2024. What Did My Car Say? Impact of Autonomous Vehicle Explanation Errors and Driving Context On Comfort, Reliance, Satisfaction, and Driving Confidence. *arXiv preprint arXiv:2409.05731* (2024).
[52] Robert Kaufman, Jean Costa, and Everlyne Kimani. 2024. Effects of multimodal explanations for autonomous driving on driving performance, cognitive load, expertise, confidence, and trust. *Scientific reports* 14 (2024).
[53] Robert Kaufman and David Kirsh. 2023. Explainable AI And Visual Reasoning: Insights From Radiology. *arXiv preprint arXiv:2304.03318* (2023).
[54] Robert Kaufman, David Kirsh, and Nadir Weibel. 2024. Developing Situational Awareness for Joint Action with Autonomous Vehicles. *arXiv preprint arXiv:2404.11800* (2024).
[55] Robert Kaufman, Emi Lee, Manas Satish Bedmutha, David Kirsh, and Nadir Weibel. 2024. Predicting Trust In Autonomous Vehicles: Modeling Young Adult Psychosocial Traits, Risk-Benefit Attitudes, And Driving Factors With Machine Learning. *arXiv preprint arXiv:2409.08980* (2024).
[56] Robert A. Kaufman, Michael Robert Haupt, and Steven W. Dow. 2022. Who's in the Crowd Matters: Cognitive Factors and Beliefs Predict Misinformation Assessment Accuracy. *Proceedings of the ACM on Human-Computer Interaction* 6 (2022), 1 – 18.
[57] Robert A Kaufman and David Kirsh. 2022. Cognitive differences in human and AI explanation. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 44.
[58] Ahmet Rıfat Kayiş, Seydi Ahmet Satici, Muhammet Fatih Yilmaz, Didem Şimşek, Esra Ceyhan, and Fuad Bakioğlu. 2016. Big five-personality trait and internet addiction: A meta-analytic review. *Computers in Human Behavior* 63 (2016), 35–40.
[59] Jonas De keersmaecker and Arne Roets. 2017. 'Fake news': Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions. *Intelligence* 65 (2017), 107–110.
[60] Jooyeon Kim, Dongkwan Kim, and Alice Oh. 2019. Homogeneity-based transmissive process to model true and false news in social networks. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 348–356.
[61] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences* 110, 15 (2013), 5802–5805.
[62] Matthew W Kreuter, David W Farrell, Laura R Olevitch, and Laura K Brennan. 2013. *Tailoring health messages: Customizing communication with computer technology*. Routledge.
[63] Kaisheng Lai, Xiling Xiong, Xiaoya Jiang, Meiqi Sun, and Lingnan He. 2020. Who falls for rumor? Influence of personality traits on false rumor belief. *Personality and Individual Differences* 152 (2020), 109520.
[64] Vivian Lai, Samuel Carton, Rajat Bhatnagar, Vera Liao, Yunfeng Zhang, Chenhao Tan, and Qingzi Vera Liao. 2022. Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (2022).
[65] Meng Zhen Larsen, Michael Robert Haupt, Tiana J McMann, Raphael E. Cuomo, and Tim Ken Mackey. 2023. The Influence of News Consumption Habits and Dispositional Traits on Trust in Medical Scientists. *International Journal of Environmental Research and Public Health* 20 (2023).
[66] M Asher Lawson and Hemant Kakkar. 2022. Of pandemics, politics, and personality: The role of conscientiousness and political ideology in the sharing of fake news. *Journal of Experimental Psychology: General* 151, 5 (2022), 1154.
[67] Jeffrey Lees, Abigail McCarter, and Dawn M Sarno. 2022. Twitter's disputed tags may be ineffective at reducing belief in fake news and only reduce intentions to share fake news among Democrats and Independents. *Journal of Online Trust and Safety* 1, 3 (2022).
[68] Jeff Levin and Matt Bradshaw. 2022. Determinants of COVID-19 skepticism and SARS-CoV-2 vaccine hesitancy: findings from a national population survey of U.S. adults. *BMC Public Health* 22 (2022).
[69] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2119–2128.
[70] Hause Lin, David G. Rand, and Gordon Pennycook. 2023. Conscientiousness does not moderate the association between political ideology and susceptibility to fake news sharing. *Journal of experimental psychology. General* (2023).

[71] Sahil Loomba, Alexandre De Figueiredo, Simon J Piatek, Kristen De Graaf, and Heidi J Larson. 2021. Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature human behaviour* 5, 3 (2021), 337–348.
[72] 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* 54 (2023), 101710.
[73] Philip M Massey, Shawn C Chiang, Meredith Rose, Regan M Murray, Madeline Rockett, Elikem Togo, Ann C Klassen, Jennifer A Manganello, and Amy E Leader. 2021. Development of personas to communicate narrative-based information about the HPV vaccine on Twitter. *Frontiers in digital health* 3 (2021), 682639.
[74] Mohsen Mosleh, Gordon Pennycook, and David G Rand. 2020. Self-reported willingness to share political news articles in online surveys correlates with actual sharing on Twitter. *Plos one* 15, 2 (2020), e0228882.
[75] Adam Mosseri. 2016. *Addressing Hoaxes and Fake News.* https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/
[76] Matthew Motta, Timothy Callaghan, and Steven Sylvester. 2018. Knowing less but presuming more: Dunning-Kruger effects and the endorsement of anti-vaccine policy attitudes. *Social Science & Medicine* 211 (2018), 274–281.
[77] Brendan Nyhan and Jason Reifler. 2010. When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior* 32 (2010), 303–330.
[78] Thomas C. O'Brien, Ryan Palmer, and Dolores Albarracín. 2021. Misplaced trust: When trust in science fosters belief in pseudoscience and the benefits of critical evaluation. *Journal of Experimental Social Psychology* 96 (2021), 104184.
[79] Ramin Okhrati and Aldo Lipani. 2021. A multilinear sampling algorithm to estimate shapley values. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 7992–7999.
[80] World Health Organization, World Health Organization, et al. 2022. Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation. 2020. *Available: web link (Accessed 2 December 2020)* (2022).
[81] Michael Pazzani, Severine Soltani, Robert Kaufman, Samson Qian, and Albert Hsiao. 2022. Expert-informed, user-centric explanations for machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12280–12286.
[82] Didem Pehlivanoglu, Tian Lin, Farha Deceus, Amber Heemskerk, Natalie C. Ebner, and Brian Stephen Cahill. 2020. The role of analytical reasoning and source credibility on the evaluation of real and fake full-length news articles. *Cognitive Research: Principles and Implications* 6 (2020).
[83] Gordon Pennycook, Adam Bear, Evan T. Collins, and David G. Rand. 2019. The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings. *Political Communication eJournal* (2019).
[84] Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general* 147, 12 (2018), 1865.
[85] Gordon Pennycook and David G Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188 (2019), 39–50.
[86] Gordon Pennycook and David G Rand. 2020. Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of personality* 88, 2 (2020), 185–200.
[87] Gordon Pennycook and David G Rand. 2021. The psychology of fake news. *Trends in cognitive sciences* 25, 5 (2021), 388–402.
[88] Andrea Pereira, Elizabeth Harris, and Jay Joseph Van Bavel. 2018. Identity concerns drive belief: The impact of partisan identity on the belief and dissemination of true and false news. *Group Processes & Intergroup Relations* 26 (2018), 24 – 47.
[89] Gennaro Pica, Antonio Pierro, Jocelyn J Bélanger, and Arie W Kruglanski. 2014. The role of need for cognitive closure in retrieval-induced forgetting and misinformation effects in eyewitness memory. *Social Cognition* 32, 4 (2014), 337–359.
[90] Michal Piksa, Karolina Noworyta, Jan Piasecki, Pawel Gwiazdzinski, Aleksander B Gundersen, Jonas Kunst, and Rafal Rygula. 2022. Cognitive processes and personality traits underlying four phenotypes of susceptibility to (mis) information. *Frontiers in Psychiatry* 13 (2022), 912397.
[91] Clara Pretus, Ali M Javeed, Diána Hughes, Kobi Hackenburg, Manos Tsakiris, Oscar Vilarroya, and Jay J Van Bavel. 2024. The Misleading count: an identity-based intervention to counter partisan misinformation sharing. *Philosophical Transactions of the Royal Society B* 379, 1897 (2024), 20230040.
[92] Clara Pretus, Camila Servin-Barthet, Elizabeth A Harris, William J Brady, Oscar Vilarroya, and Jay J Van Bavel. 2023. The role of political devotion in sharing partisan misinformation and resistance to fact-checking. *Journal of Experimental Psychology: General* (2023).
[93] Man pui Sally Chan and Dolores Albarracín. 2023. A meta-analysis of correction effects in science-relevant misinformation. *Nature Human Behaviour* 7 (2023), 1514 – 1525.

[94] Beatrice Rammstedt and Oliver P John. 2007. Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of research in Personality* 41, 1 (2007), 203–212.

[95] Arne Roets et al. 2017. 'Fake news': Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions. *Intelligence* 65 (2017), 107–110.

[96] Arne Roets and Alain Van Hiel. 2011. Item selection and validation of a brief, 15-item version of the Need for Closure Scale. *Personality and individual differences* 50, 1 (2011), 90–94.

[97] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. Can the crowd identify misinformation objectively? The effects of judgment scale and assessor's background. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*. 439–448.

[98] Kevin Roitero, Michael Soprano, Beatrice Portelli, Damiano Spina, Vincenzo Della Mea, Giuseppe Serra, Stefano Mizzaro, and Gianluca Demartini. 2020. The covid-19 infodemic: Can the crowd judge recent misinformation objectively?. In *Proceedings of the 29th ACM international conference on information & knowledge management*. 1305–1314.

[99] Peter A Rosen and Donald H Kluemper. 2008. The impact of the big five personality traits on the acceptance of social networking website. *AMCIS 2008 proceedings* (2008), 274.

[100] Emily Saltz, Soubhik Barari, Claire Leibowicz, and Claire Wardle. 2021. Misinformation interventions are common, divisive, and poorly understood. *Harvard Kennedy School Misinformation Review* (2021).

[101] 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*. 1–6.

[102] Brinda Sampat and Sahil Raj. 2022. Fake or real news? Understanding the gratifications and personality traits of individuals sharing fake news on social media platforms. *Aslib Journal of Information Management* 74, 5 (2022), 840–876.

[103] Johanes Schneider and Joshua Handali. 2019. Personalized explanation in machine learning: A conceptualization. *arXiv preprint arXiv:1901.00770* (2019).

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

[105] Donghee Shin. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies* 146 (2021), 102551.

[106] Mohit Singhal, Chen Ling, Nihal Kumarswamy, Gianluca Stringhini, and Shirin Nilizadeh. 2022. SoK: Content Moderation in Social Media, from Guidelines to Enforcement, and Research to Practice. *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)* (2022), 868–895.

[107] Matthew Smith and Francisco Alvarez. 2021. Identifying mortality factors from Machine Learning using Shapley values–a case of COVID19. *Expert Systems with Applications* 176 (2021), 114832.

[108] Severine Soltani, Robert A Kaufman, and Michael J Pazzani. 2022. User-centric enhancements to explainable ai algorithms for image classification. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 44.

[109] Yanqing Sun and Juan Xie. 2024. Who shares misinformation on social media? A meta-analysis of individual traits related to misinformation sharing. *Computers in Human Behavior* 158 (2024), 108271.

[110] Gunnvald B Svendsen, Jan-Are K Johnsen, Live Almås-Sørensen, and Joar Vittersø. 2013. Personality and technology acceptance: the influence of personality factors on the core constructs of the Technology Acceptance Model. *Behaviour & Information Technology* 32, 4 (2013), 323–334.

[111] Viren Swami, Ulrich S Tran, Stefan Stieger, Jakob Pietschnig, Ingo W Nader, and Martin Voracek. 2016. Who believes in the giant skeleton myth? An examination of individual difference correlates. *SAGE Open* 6, 1 (2016), 2158244015623592.

[112] Edson C Tandoc Jr, Zheng Wei Lim, and Richard Ling. 2018. Defining "fake news" A typology of scholarly definitions. *Digital journalism* 6, 2 (2018), 137–153.

[113] Kathie M d'I Treen, Hywel TP Williams, and Saffron J O'Neill. 2020. Online misinformation about climate change. *Wiley Interdisciplinary Reviews: Climate Change* 11, 5 (2020), e665.

[114] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *science* 359, 6380 (2018), 1146–1151.

[115] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[116] Sai Wang. 2021. Moderating uncivil user comments by humans or machines? The effects of moderation agent on perceptions of bias and credibility in news content. *Digital journalism* 9, 1 (2021), 64–83.

[117] Tianjiao (Grace) Wang and Wenting Yu. 2022. Alternative sources use and misinformation exposure and susceptibility: The curvilinear moderation effects of socioeconomic status. *Telematics Informatics* 70 (2022), 101819.

[118] Przemyslaw M Waszak, Wioleta Kasprzycka-Waszak, and Alicja Kubanek. 2018. The spread of medical fake news in social media–the pilot quantitative study. *Health policy and technology* 7, 2 (2018), 115–118.

[119] Donna M Webster and Arie W Kruglanski. 1994. Individual differences in need for cognitive closure. *Journal of personality and social psychology* 67, 6 (1994), 1049.

[120] Thomas Wood and Ethan Porter. 2019. The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior* 41 (2019), 135–163.

[121] Eddy SF Yeung and Kai Quek. 2024. Self-reported political ideology. *Political Science Research and Methods* (2024), 1–22.

A Additional Tables

A.1 Mean Comparison of MTurk and SONA

Based on Table 4, the MTurk sample is older and more conservative compared to SONA. MTurkers also show a statistically significant difference with higher Trust in News Outlets, Trust in Religious Leaders, Trust in Elected Officials, and NFCC compared to SONA respondents. MTurkers also have slightly more favorable attitudes towards warning tags on average than SONA.

Table 4. Mean Comparisons in Factor Measures by Platform

	MTurk	**SONA**	**Diff**	***Scale Range***
Age	**37.25**	**21.63**	**15.62**	Open
Political Conservativeness	**3.39**	**2.54**	**0.85**	(1 to 6)
Trust News	**3.09**	**2.04**	**1.05**	(1 to 4)
Trust Medical	3.07	3.13	-0.06	(1 to 4)
Trust Religious	**2.69**	**1.65**	**1.04**	(1 to 4)
Trust Elected	**2.85**	**1.83**	**1.02**	(1 to 4)
NFCC	**4.29**	**3.97**	**0.32**	(1 to 6)
CRT	0.98	1.15	-0.17	(0 to 3)
Conscientious	3.34	3.44	-0.10	(1 to 5)
Extraversion	NA	3.08	NA	(1 to 5)
Openness	NA	3.74	NA	(1 to 5)
Neuroticism	NA	3.12	NA	(1 to 5)
Agreeableness	NA	3.46	NA	(1 to 5)
Aggregate Tag Score	**3.60**	**3.46**	**-0.1**	(1 to 5)

Note: Emboldened characteristics are significantly different at p < 0.05 in the two-sided test of equality for column means.

A.2 Bivariate Correlations for SONA Sample Only

Bivariate correlations were run on the SONA sample between tag attitudes items and individual traits as shown in Table 5. Positive correlations were shown between the composite tag attitude scale and BFI traits for Openness (r = 0.40, p < 0.01), Conscientiousness (r = 0.19, p < 0.05), and Agreeableness (r = 0.29, p < 0.01). For demographic variables, the tag attitude scale was positively correlated with stronger identification with being a Democrat (0.36, p < 0.01) and higher Trust in Medical Scientists (r = 0.27, p < 0.01). However, Trust in Religious Leaders was negatively correlated with tag attitudes (r = -0.20, p <.01). All variables that showed significant correlations with the composite attitude scale also showed correlations with multiple individual tag items. While CRT shows no significant correlation with the tag attitude scale in the SONA sample, it was positively associated with the tag item "I am more likely to research the post's topic in order to evaluate the content myself" (r = 0.18, p < 0.05).

A.3 Multiple Regression Results for SONA Sample Only

Multiple regression model results with Shapley coefficients for the SONA sample are shown in Table 6. The current model is shown to be a strong predictor of tag attitude variance (R^2 of 0.38 and adjusted R^2 of 0.30). Political orientation is the most influential variable explaining 39.01% of variance for tag attitudes as shown from the Shapley results. Further, political liberalism was

Table 5. Bivariate correlations between tested variables and tag attitude items (SONA)

	Overall	"I usually ignore the tag"	"I am less likely to read the post"	"I am less likely to engage with the post through sharing, replying, or liking"	"I evaluate the post content based on who the author is"	"I am more likely to research the post's topic in order to evaluate the content myself"	"I approach the post content with suspicion or hesitancy"	"I view the author of the post in a negative way"	"I trust that the post is, in fact, misinformation"	"I appreciate the social media platform's attempt to warn me about potential misinformation"
CRT	0.070	0.020	-0.036	-0.074	0.130	**0.180***	0.111	0.010	0.000	0.065
NFCC	0.050	0.001	0.096	0.0270	-0.025	-0.030	0.063	0.050	-0.060	0.0280
Openness	**0.400****	**-0.243****	0.091	**0.244****	**0.207****	**0.280****	**0.310****	0.078	0.130	**0.359****
Conscientious	**0.190***	**-0.199****	0.049	0.100	0.086	-0.007	**0.245****	-0.020	0.060	**0.228****
Agreeableness	**0.290****	-0.125	0.139	**0.173***	**0.194****	0.061	**0.268****	0.066	0.080	**0.256****
Extroversion	0.110	-0.111	-0.003	0.130	0.110	0.021	0.006	-0.081	0.097	0.101
Neuroticism	0.070	-0.013	0.088	-0.105	0.032	0.039	-0.019	0.067	0.009	-0.040
Politics	**-0.360****	**0.285****	**-0.194***	**-0.230****	-0.005	-0.125	**-0.254****	**-0.300****	**-0.332****	**-0.415****
Trust Elected	-0.020	0.139	0.110	-0.024	0.068	-0.001	-0.140	0.160	-0.007	-0.111
Trust News Media	-0.030	0.053	0.130	0.019	-0.088	0.072	-0.030	-0.043	-0.025	-0.004
Trust Med Sci	**0.270****	-0.102	0.125	**0.265****	0.129	**0.252****	**0.339****	0.043	0.040	**0.251****
Trust Religious	**-0.200****	0.127	-0.144	-0.114	-0.015	-0.094	**-0.188***	**-0.190***	-0.061	-0.162
Age	-0.050	0.067	-0.051	-0.028	-0.092	-0.043	-0.114	0.104	-0.006	-0.082

Attitude responses prefaced with "When I see a social media post tagged as potential misinformation..."
Significance codes: $^{*}p < 0.05$, $^{**}p < 0.01$

Table 6. Multiple regression – Tag attitude aggregate as dependent variable (SONA)

	β **coef**	**Shapley** R^2	**std err**	**t-value**	**p-value**
Politics	**-0.28**	**39.01%**	**0.07**	**-4.05**	**0.00**
CRT	-0.01	0.30%	0.04	-0.34	0.73
NFCC	0.09	1.95%	0.09	1.08	0.28
Openness	**0.21**	**19.18%**	**0.10**	**2.00**	**0.05**
Conscientiousness	0.15	5.61%	0.09	1.55	0.12
Agreeableness	0.08	7.95%	0.11	0.77	0.44
Extraversion	-0.03	0.91%	0.10	-0.32	0.75
Neuroticism	0.09	2.72%	0.08	1.21	0.23
Trust Elected	0.07	0.86%	0.08	0.94	0.35
Trust News Media	-0.01	0.34%	0.07	-0.11	0.91
Trust Med Sci	**0.15**	**13.54%**	**0.07**	**2.24**	**0.03**
Trust Religious	-0.07	7.21%	0.07	-1.00	0.32
Age	0.00	0.42%	0.01	0.02	0.98
R^2	0.38				
Adj R^2	0.30				

positively associated with tag attitudes when controlling for all other tested variables (β = -0.28, p < 0.001). BFI Openness was the second most influential variable contributing up to 19.18% of tag attitude variance, and was shown to be positively associated when controlled for all other factors (β = 0.21, p < 0.05). Trust in Medical Scientists was the third most influential variable contributing to 13.54% of tag attitude variance. This model also shows that Trust in Medical Scientists is positively correlated with tag attitudes when controlled for all other traits and demographic factors (β = 0.15, p < 0.05). No other variables in the model showed a significant effect with aggregate tag attitude.

A.4 Bivariate Correlations for MTurk Sample Only

Table 7 shows bivariate correlations for the MTurk sample. NFCC was positively correlated with the tag attitude scale (r = 0.447, p < 0.01). Further, Conscientiousness within the MTurk sample was negatively correlated with the tag attitude scale (r = -0.161, p < 0.05). All remaining variables except for Trust in Elected Officials showed no statistically significant effects with the composite tag attitude scale. While CRT was not statistically significant with respect to aggregate tag attitudes, those higher in CRT were less likely to agree with the individual statement "I usually ignore the tag" (r = -0.230, p < 0.01). Political orientation did not show a significant correlation with the aggregate attitude scale, however, being liberal was correlated with agreeing with the item "I am less likely to read the post" (r = -0.165, p < 0.05) when seeing a tag for potential misinformation. For trust in new media, only "I appreciate the social media platform's attempt to warn me about potential misinformation" showed a significant effect (r = 0.182, p < 0.05). Trust in Medical Scientists had three significant correlations with "I usually ignore the tag", "I evaluate the post content based on who the author is", and "I appreciate the social media platform's attempt to warn me about potential misinformation" despite the aggregate not being significant. Trust in Religious Leaders had significant correlations with "I usually ignore the tag" and "I am more likely to research the post's topic in order to evaluate the content myself". Lastly, the response "I usually ignore the tag" was negatively correlated with age (r = -0.154, p < 0.05).

Table 7. Bivariate correlations between tested variables and tag attitude items (MTurk)

	Overall	"I usually ignore the tag"	"I am less likely to read the post"	"I am less likely to engage with the post through sharing, replying, or liking"	"I evaluate the post content based on who the author is"	"I am more likely to research the post's topic in order to evaluate the content myself"	"I approach the post content with suspicion or hesitancy"	"I view the author of the post in a negative way"	"I trust that the post is, in fact, misinformation"	"I appreciate the social media platform's attempt to warn me about potential misinformation"
CRT	-0.005	**-0.230****	0.031	0.013	0.001	-0.115	0.091	-0.038	-0.134	-0.054
NFCC	**0.447****	**0.399****	**0.218****	**0.320****	**0.272****	**0.145***	**0.297****	**0.412****	**0.259****	**0.285****
Conscientious	**-0.161***	**-0.214****	-0.125	-0.127	-0.061	-0.041	-0.033	**-0.167***	**-0.180****	-0.083
Politics	-0.079	0.122	**-0.165***	-0.028	0.048	0.005	-0.101	-0.019	-0.095	0.003
Trust Elected	0.084	0.133	-0.005	-0.053	0.117	0.123	0.085	0.053	0.096	0.124
Trust News Media	0.075	-0.017	0.020	0.009	0.138	0.035	0.029	0.015	0.010	**0.182***
Trust Med Sci	0.109	**-0.160***	-0.061	0.146	**0.202****	0.128	0.137	-0.080	0.053	**0.206****
Trust Religious	0.015	**0.212****	-0.066	-0.057	0.088	**0.162***	0.069	-0.054	0.013	0.100
Age	-0.077	**-0.154***	-0.114	-0.108	0.032	-0.019	-0.087	-0.016	-0.129	-0.036

Attitude responses prefaced with "When I see a social media post tagged as potential misinformation..."

Significance codes: $^{*}p < 0.05$, $^{**}p < 0.01$

Table 8. Multiple regression – Tag attitude aggregate as dependent variable (MTurk)

	β coef	**Shapley R^2**	**std err**	**t-value**	**p-value**
Politics	**-0.05**	**7.78%**	**0.02**	**-2.35**	**0.02**
CRT	0.06	2.01%	0.04	1.54	0.13
NFCC	**0.29**	**59.77%**	**0.04**	**6.59**	**0.00**
Conscientiousness	**-0.18**	**9.18%**	**0.07**	**-2.49**	**0.01**
Trust Elected	0.02	1.46%	0.05	0.42	0.68
Trust News Media	0.08	4.99%	0.05	1.74	0.08
Trust Med Sci	**0.11**	**9.94%**	**0.04**	**2.52**	**0.01**
Trust Religious	**-0.09**	**3.80%**	**0.04**	**-1.99**	**0.05**
Age	0.00	1.07%	0.00	-0.62	0.53
R^2	0.37				
Adj R^2	0.32				

A.5 Multiple Regression Results for MTurk Sample Only

Regression results for the MTurk sample are shown in Table 8. Within this model, NFCC is the most influential variable explaining up to 59.77% of variance for the tag attitude scale. Further, NFCC is positively correlated with favorable tag attitudes and is statistically significant when controlled for all other factors (β = 0.29, p < 0.001). The second most influential variable, Trust in Medical Scientists, is also positively correlated with tag attitudes (β = 0.11, p<.01) and explains 9.94% of model variance. Conscientiousness, conservatism, and trust in religious figures are all negatively correlated with tag attitudes (β = -0.18, p < 0.01;β = -0.05, p < 0.05; = -0.09, p < 0.05). The total R^2 of the model is 0.37 and the adjusted R^2 is 0.32.

Received July 2024; revised December 2024; accepted March 2025