scientific reports

Check for updates

OPEN

People adhere to content warning labels even when they are wrong due to ecologically rational adaptations

Benjamin D. Horne[1]✉ & Dorit Nevo[2]

In this paper, we build on the theory of ecologically rational heuristics to demonstrate the effect of erroneously placed warning labels on news headlines. Through three between-subjects experiments (n = 1313), we show that people rely on warning labels when choosing to trust news, *even when those labels are wrong*. We argue that this over-reliance on content labels is due to ecological rationality adaptations to current media environments, where warning labels are human-generated and mostly correct. Specifically, news consumers form heuristics based on past interactions with warning labels, and those heuristics can spill-over into new media environments where warning label errors are more frequent. The most important implication of these results is that it is more important to thoughtfully consider what information needs to be labeled than it is to attempt to label all false information. We discuss how this implication impacts our ability to scale warning label systems.

The prevalence of online disinformation and misinformation has driven scientists and media platforms to prioritize the reduction of their impact[1]. One commonly adopted solution is to attach warning labels to questionable content, typically determined by external fact-checkers[2,3]. Warning labels from third-party fact checkers have been used on multiple social media platforms, including Twitter/X, Facebook, Instagram, and TikTok. Studies so far have shown that these warning labels are effective at reducing the sharing of false information, trust in false information, and the accuracy perception of false information[3–5]. Further, this effectiveness holds relatively well across political party lines[6–8] and other individual differences[3].

However, it has been argued that warning label effectiveness is limited by its scalability[3]. Current warning label systems rely on human labor and careful deliberation to ensure warnings are correct. For example, on Facebook when questionable content is identified by platform users, it is then reviewed by independent fact checkers. This review may involve multiple, time-consuming steps, such as "calling sources, consulting public data, authenticating videos and images and more"[9]. This careful process faces a significant scalability challenge due to the substantial amount of content uploaded to media platforms daily (e.g. see Ref.[10]). One proposed method to increase the scalability of warning label systems – and the coverage of false information labeled—is to train Artificial Intelligence (AI) to label content[11]. Many of these proposed systems have shown high accuracy in lab settings[12–14], and label experiments suggest that warnings from AI are effective[4,5,8,15], although perceptions of AI in content moderation vary[16].

There is a concern that with increased use of automation, so too will the frequency of labeling errors increase[17,18]. We can conceive of two types of errors: (1) if a warning label is erroneously placed on *true* information (false positive), and (2) if a warning label is *not* placed on false information (false negative). For example, if a warning label is erroneously placed on true information (for example, about the safety of vaccines), this warning may push consumers to wrong inference (for example, not get vaccinated towards a deadly disease) or to distrust related information in the future. On the other hand, if a warning label is *not* placed on *false* information (about the safety of vaccines), the same outcomes may occur.

Prior literature further suggests that erroneously placed warning labels can lead individuals to discard true information and may reduce the perceived credibility of such information[19]. This work and related work on memory during eyewitness testimony provides some evidence that warning labels placed on true news can have an adverse effect on correct memories when recalling events (*tainted truth effect*)[19,20]. Inversely, some experiments show that false headlines that fail to get labeled may be deemed validated and consequently seen as more accurate (*implied truth effect*)[21]. As put by Freeze et al.[19]: "Even legitimate misinformation warnings, if not

[1]School of Information Sciences, University of Tennessee Knoxville, Knoxville, Tennessee, USA. [2]Lally School of Management, Rensselaer Polytechnic Institute, Troy, New York, USA. ✉email: bhorne6@utk.edu

fully deployed, can enhance the effects of misinformation in the larger system." Together, these works illustrate the potential dangers of both false positive and false negative errors within information systems that deploy content warning labels.

In this paper we focus on erroneously placed warning labels on news headlines. While the abovementioned literature studied the potential outcomes of such errors (e.g. the tainted truth and implied truth effects), the impact of content label mistakes on information trust requires further investigation. Specifically, we ask whether, and to what extent, people base their news veracity judgements on warning labels across different underlying information veracities. In other words, do people over-rely on content warning labels? And if so, what explains this over-reliance? We ask these questions within the hypothetical context of social media platforms moving away from employing third-party fact checkers to deploying AI tools to label content instead. While this context is the motivating backdrop for our study, its results have implications well beyond this specific context.

In a series of three online experiments, we find that warning labels significantly decrease trust in the information they are attached to, no matter the veracity of the underlying information. We argue that this effect is due to information consumers making ecologically rational adaptations to current media environments, which spill-over into new media environments where warning label mistakes are more frequent. However, we also show that with relatively little experience with the new media environment, information consumers can calibrate their decision heuristics to perform better and not over-rely on the erroneously placed labels. Paradoxically, this result implies that the harm from *infrequent errors* may be as bad as the harm from a less accurate system with more errors. This *rare failure effect* suggests that we should further study the magnitude of news veracity judgement in different contexts and carefully consider what information should have a warning label rather than sacrifice accuracy for information coverage. To explore this effect further, we conclude with a study that examines news consumers' sensitivity to warning labels under various conditions.

Background

Dual process theories of information, such as the heuristic-systematic model (HSM[22]), posit two paths that are involved in information processing: one through the use of heuristics and the other through deeper processing of the information content. The heuristic route involves the use of mental shortcuts to make judgements in information processing, whereas the systematic route involves deliberate and careful processing of the information, which requires people to engage in analytical thinking about the information provided[22]. The model posits that heuristic and systematic processing are not mutually exclusive and can co-exist[23]. To determine the extent of information processing that a person may engage in, the model proposes that people try to be as efficient as possible, using the less effortful mode of processing when possible[23], and that people may consider additional effort when they experience a gap between their actual and desired confidence in the information[24]. The bigger the gap, the more likely a person is to engage in systematic processing, trading-off effort with confidence[25].

One assumption in the study of false information is that consumers will default to cognitive biases (such as confirmation bias[26]) and personal heuristics in processing information[27]. Indeed, a significant amount of prior work has shown that attributes of the information itself and how that information is perceived by the consumer can significantly impact information trust. For example, trust may change based on the source of the information, the coherency/familiarity/fluency of the information, the writing style, or the information's alignment with prior beliefs[28]. More recent studies suggest that warning label interventions are also commonly used heuristics that can be broadly effective in shifting information trust and sharing behaviors[3]. In fact, it may be that relying on warning labels has become the default decision heuristic rather than other information-based heuristics or cognitive biases because of their ecological rationality.

Ecologically rational heuristics are a subset of heuristics that are effective within a particular environment. The study of ecological rationality examines the environmental structures that make such heuristics under conditions of uncertainty[29]. Specifically, the ecological rationality perspective emphasizes the fit between the heuristic chosen and the requirements of the task, and it seeks those situations in which the heuristic performs better than more complex strategies[30]. For example, the recognition heuristic can be considered an ecologically rational heuristic (ERH) if it leads to correct decisions over half of the time within a particular environment[29]. Similarly, warning labels can be considered ecologically rational under specific media environments if they consistently lead to correct judgements of news veracity.

Building on the notion of ERH, we can examine the effect of rational adaptation to our media environments on trust in warning labels. To the best of our knowledge, *currently deployed* human-labor warning labels systems are quite successful, with errors being somewhat rare, although they do occur[31]. Hence, if adherence to content warning labels frequently leads to correct decisions in our current media systems, information consumers may form ERH using those labels. In other words, the heuristic "trust a warning label" may provide the desired confidence in the information with reasonably low effort. Hence, it would be considered an ERH; utilizing this heuristic is expected to result in better outcomes than employing more complex decision strategies or other heuristics. Note that this does not preclude information consumers from engaging in systematic processing of the information if deemed necessary for sufficient confidence in the information. It simply allows them to exert less effort in processing the information through the heuristic processing path.

However, as stated by Todd and Gigerenzer[29], the environment plays an important role in ecological rationality: "individuals can be led to use particular heuristics in inappropriate environments and consequently make errors…". This may be the case of applying the current ERH formed based on human generated warning labels to content warning labels that are generated by algorithms, or this may be the case of applying the current ERH to warning labels on a new social media platform with different content moderation rules and norms (a new media environment).

Algorithmically generated warning labels are similar – but not identical—to human generated fact checking labels. As such, they represent a new environment in the context of ecological rationality. Algorithms to generate warning labels range widely in terms of technical methods used, from Machine Learning models using the content in news articles and claims[4,32], to network-based models for news outlets and social media accounts[33], to knowledge-graph models for fact-checking[34], and even Large Language Models (LLMs) for outlet-level credibility predictions[35,36]. Across these technical methods, many classifiers have been proposed with varying levels of accuracy and robustness. For example, Shu et al.[12] demonstrates a tool that can classify social media posts into true and false groups with between 80 and 90% accuracy. Similar levels of accuracy have been achieved by other tools with different underlying designs[13,37]. Importantly, these metrics come from limited test settings and are unlikely to translate perfectly into real-life systems[17,18].

Of the studies that have experimentally examined AI warning labels, they have shown that while AI warnings labels are consistently effective compared to having no warning labels, their relative efficacy compared to other types of warning labels varies. For example, Yaqub et al.[15] showed that labels from third-party fact checkers deterred the sharing of false news stories more than warnings from AI. On the other hand, Horne[8] showed that AI warning labels decreased trust in false information modestly more than warning labels from third-party fact checkers, particularly for people who do not trust news organizations. Given these mixed results we suspect that people would transfer their reliance on ERH developed in the human fact-checking environment to the algorithmic one. However, due to the variability in accuracy and coverage from human fact-checkers to algorithmic models or even between different algorithms, what once was ecologically rational may no longer be so.

Hence, in a series of three studies, we explore the following research questions: **(RQ1)** do people rely on warning labels as heuristics in forming judgement on news veracity, even when those labels are incorrect? And **(RQ2)** If people indeed continue to rely on erroneously placed labels, can ecological rationality explain this over-reliance? We present the empirical studies first and discuss the important implications of this study in the following sections.

Empirical studies: trust in news with attached warning labels

We designed three experiments to explore the impact of erroneously placed warning labels on trust in information, and whether rational adaptations to current media environments explain the errors made by participants. Our design in all three studies is a 2×2 online experiment where warning labels are either correctly placed on false news and not placed on true news, versus falsely placed on true news and not placed on false news. This design is illustrated in Table 1.

The study task, under all conditions, was the same: observe a news headline, with or without a warning label, and indicate the extent to which you trust this news headline, where trust is a uni-dimensional conceptualization ranging from strongly distrust to strongly trust. We chose a uni-dimensional conceptualization trust to align with prior warning label studies that examined trust (e.g. Ref.[8]).

Study 1 was meant to explore our suspected spill-over of heuristics from current media environments to the experimental setup. In this study, we did not provide any "training" with the warning label system but asked respondents to "jump into the task" immediately. Hence, theoretically, their experience with warning labels in real-life should spill-over into how they interact with warning labels in the experiment. If warning labels were indeed the default heuristic over other forms of information processing, the participants should adhere to the label advice in the experiment, no matter the information it is attached to.

Study 2 delve deeper into this spill-over effect by first presenting participants with a habituation phase in which the accuracy of the warning labels was manipulated and then asking them to evaluate the veracity of different news headlines, with or without warning labels. In this second study we allowed respondents to interact with the warning labels system for an extended period before we evaluated their trust in the headlines provided. The accuracy of the warning labels system participants interacted with differed under the different study conditions, thereby creating two different environments that potentially changed their overreliance or aversion to following the warning labels advice. Study 2 further demonstrates how ERH can be changed through habituation.

Finally, as previously reviewed, ecological rationality seeks those heuristics that perform better than more complex strategies. In study 3 we focused on examining the performance of heuristics using Signal Detection Analysis. Specifically, we examined an environment in which labels were placed on *all* the content (both true and false) and were framed as either warning or supporting messages. We then computed respondents' ability to correctly detect false news in this labeling environment. We explain each study in detail below.

Study 1: Do labeling errors influence trust in true information?

In this first study, we wanted to better understand if warning label errors matter in immediate information trust decisions. To this end, we conducted a three-arm between-subjects experiment, in which participants ($n = 366$) rated their trust in true and false headlines (10 headlines per participant, making 3,660 headline-level data points). Participants were recruited for the study on the survey platform Prolific – which has been shown

	Warning	No warning
False headline	Correct	False negative (cost: trusting in false news)
True headline	False positive (cost: distrusting in true news)	Correct

Table 1. Types of trust errors by consumers when warning labels are correctly or incorrectly placed.

to produce high quality data[38]—and data were collected in October 2023. Participants came from a standard sample of U.S. residents. In terms of demographics, about 56% identified as female and 43% as male, 72% of respondents were white, and the average age of participants was 43.35 (median 40).

Participants were assigned to one of three conditions: 1. No warning labels (control), 2. Correctly placed warning labels, and 3. Erroneously placed warning labels. In the *Correct* condition, 100% of the false headlines had warning labels attached and 0% of the true headlines had warning labels attached. In the *Error* condition, the inverse happened: 100% of the true headlines had warning labels attached and 0% of the false headlines had warning labels attached. To ensure the headlines were fresh[39], the false headlines used in this study were randomly selected from recently fact-checked headlines/social media posts from Snopes, PolitiFact, FactCheck.org, or Reuters Fact Check. The true headlines in this study were randomly sampled from both recently fact-checked headlines and from recent articles published by Reuters, the Associated Press, or NPR – outlets that are generally considered reliable and centrist. Approximately half of the true headlines were from fact-checkers and half from Reuters, the Associated Press, or NPR. This choice ensures that our true headlines were a mixture of what one would consider typical true headlines and true headlines that had to be fact-checked. The headlines could be from any topic, not limited to politics or health. All warning labels attached to these headlines read: "Warning: An AI Tool Says This Story is False". An example of a true headline used in the study with and without a label attached is shown in Fig. 1. The warning label design was simple and roughly followed the design used in prior warning label studies (for example Refs.[4,8,40]). The source of the news headline was not shown across all conditions. Participants were approximately balanced across the three conditions, with 100 in control, 123 in the correct label condition, and 116 in the error label condition.

Since this experiment captured repeated measurements for each participant, our primary method of analysis was a mixed-effects regression with random intercepts (also called a multilevel model), where trust scores for individual headlines (level 1) are clustered under participants (level 2). We fit two models: one for trust in false headlines and another for trust in true headlines. In these models, the dependent variable was post–level trust and the independent variables were the condition a participant was assigned to. We chose to separate the headline veracity into two models for ease of interpretation across the conditions and headline veracity, although these models could be combined by using headline veracity as random slopes in the model (e.g. Ref.[41]) or by creating a single discernment variable similar to what other studies in this area have done (see for example Refs.[42,43]). Post-level trust was measured on a 5-point Likert scale (distrust completely, somewhat distrust, neither trust nor distrust, somewhat trust, and trust completely). In the regression, we entered responses as a number between −1 and 1 that directly mapped to the Likert–scale used in the survey, where −1 is "Distrust Completely" and 1 is "Trust Completely", creating negative scores for distrust, zero for uncertain, and positive scores for trust.

We also calculated another score called participant-level trust, which is a number between −5 and 5, where −5 means they completely distrusted all false (true) posts and 5 means they completely trusted all false (true) posts. Specifically, for each false (true) headline, if a participant completely distrusted, we subtract 1, if they somewhat distrusted, we subtract 0.5, if they somewhat trusted, we add 0.5, and if they completely trusted, we add 1. Note, this score is calculated per ground truth, hence there are two scores: participant-level trust in false headlines and participant-level trust in true headlines. This score is like the participant level-scoring used by Horne[8]. We used this score to further describe the data, not for modeling.

Study 1 results

In Table 2, we show the results of two mixed-effects regression models: one for trust in true headlines and one for trust in false headlines. In Fig. 2, we show the distribution changes of both post-level trust and participant-level trust between each condition and headline ground truth.

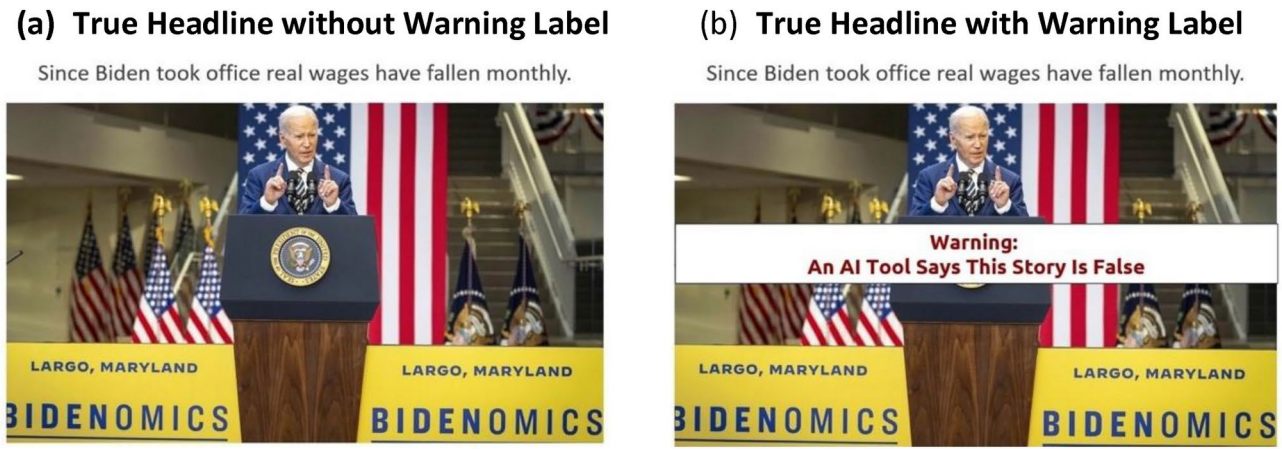

Fig. 1. Example of headline with and without warning label. This headline is a claim made by the West Virginia Republican Party that was fact-checked as true by PolitiFact. While this headline is political in topic, this was not a requirement in selecting headlines.

	(a) Trust in true headlines				(b) Trust in false headlines			
	Coef.	Std.err	[0.025	0.975]	Coef.	Std.err	[0.025	0.975]
Intercept	− 0.041	0.034	− 0.025	0.107	− **0.206****	0.035	− 0.275	− 0.138
Correct labeling	− 0.022	0.046	− 0.111	0.067	− **0.168*****	0.047	− 0.261	− 0.076
Error labeling	− **0.298*****	0.046	− 0.389	− 0.208	0.008	0.048	− 0.085	0.101
Participant	0.045	0.017	–	–	0.070	0.021	–	–

Table 2. Multilevel regression models with random intercepts where posts are clustered under participants. The dependent variable is post-level trust (ranging from − 1 to 1). Each condition is in reference to the control condition where no labels are presented. ***p < 0.001, **p < 0.01, *p < 0.05. Significant values are in bold.

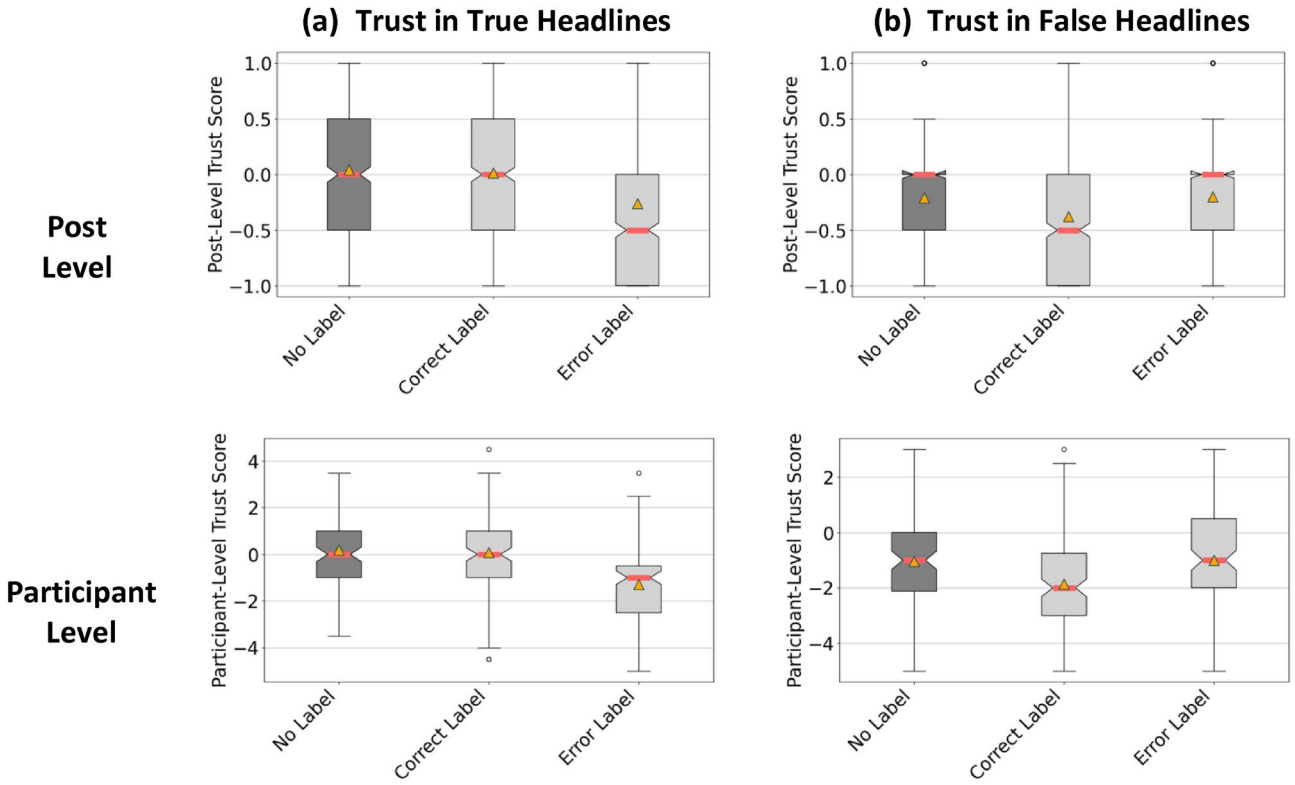

Fig. 2. Post-level and participant-level trust scores across (**a**) true headlines and (**b**) false headlines, where lower on the Y-axis is less trust. Note that false information is not labeled in the mistake label condition, while true information is not labeled in the correct label condition.

The key takeaway from this experiment is that warning labels significantly influenced trust decisions no matter the underlying veracity of the information they were attached to. As shown in Fig. 2b and Table 2b, when participants were presented with false headlines that had warning labels attached, trust in those headlines significantly decreased compared to the control ($p < 0.001$). This result is expected based on prior warning label studies[3]. However, as shown in Fig. 2a and Table 2a, this effect was also present when those headlines were true rather than false. That is, when participants were presented with true headlines that had warning labels attached, trust in those headlines significantly decreased compared to the control ($p < 0.001$). In fact, as shown by the coefficients in Table 2, and supported by calculating the Cohen's effect sizes, the effect of placing a warning label on a true headline was nearly double that of not placing a warning label on a false headline. This may imply that the presence of warning labels invokes heuristic processing that is based on the label itself. It may also imply that people are more likely to doubt themselves when choosing to trust the true headlines with labels. In study 3, we will explore these differences further. In both cases, trust in headlines without labels did not significantly differ from control.

Study 2: Do ecologically rational adaptations explain adherence to erroneous warning labels?

In study 1 we examined the spill-over of ERH from our current media environment to an algorithmic one. To further study this effect in a more controlled experiment, and to understand how the effect can be enhanced or mitigated, we conducted a second study. This study mirrored the first in its basic design and task, but we

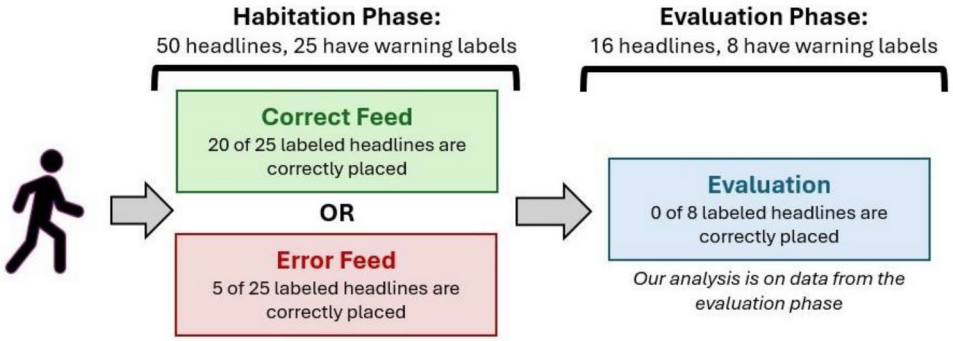

Fig. 3. Flow chart of study 2. In the *CorrectFeed*, participants saw 50 headlines, where 25 were unlabeled, 20 false headlines were correctly labeled, and 5 true headlines were labeled with erroneous warnings. In the *ErrorFeed*, participants saw the same 50 headlines, where 25 were unlabeled, 5 false headlines were correctly labeled, and 20 true headlines were labeled with erroneous warnings. All participants saw the same evaluation phase. In the evaluation phase, participants saw 16 headlines, where 8 true headlines were labeled with erroneous warnings, 8 false headlines were unlabeled.

	(a) Trust in true headlines (in the evaluation phase)				(b) Trust in false headlines (in the evaluation phase)			
	Coef.	Std.Err	[0.025	0.975]	Coef.	Std.Err	[0.025	0.975]
Intercept	− **0.206*****	0.023	− 0.251	− 0.162	− **0.141*****	0.020	0.180	− 0.102
CorrectFeed habituation	− **0.089****	0.032	− 0.152	− 0.025	0.006	0.028	− 0.049	0.061
Participant	0.098	0.016	–	–	0.063	0.012	–	–

Table 3. Multilevel regression models with random intercepts where posts (Level 1) are clustered under participants (Level 2). The dependent variable is post-level trust (ranging from -1 to 1). The *CorrectFeed* condition is in reference to the *ErrorFeed* condition. ***p < 0.001, **p < 0.01, *p < 0.05. Significant values are in bold.

added a habituation phase preceding the evaluation phase. Specifically, we designed an experiment inspired by Orchinik et al.[44], in which participants (n = 508) indicated their trust in headlines across the two phases. In the habituation phase, participants were assigned to one of two conditions: the *CorrectFeed* condition and the *ErrorFeed* condition. In the CorrectFeed condition, 80% of the labeled headlines had correctly placed warning labels. In the ErrorFeed condition, only 20% of the labeled headlines had correctly placed warning labels. Both feeds contained the same set of 50 headlines, balanced between true and false, shown in random order. In both feeds, 25 headlines had warning labels. After completing the habituation phase, participants all entered the same evaluation phase with 8 false headlines and 8 true headlines, in which 100% of the warning labels were *incorrectly placed* (replicating the *Error* condition from study 1). Participants were not notified of the change from habitation phase to evaluation phase, allowing us to test if adjustments in warning label adherence spill over into the new media environment. Again, these headlines were shown in random order and were randomly selected from either recently fact-checked headlines/social media posts or articles from Reuters, the Associated Press, or NPR. A flow chart of this study design can be found in Fig. 3.

The study was launched on Prolific and completed in June 2024. In terms of demographics, about 48% identified as female and 46% as male, 60% of respondents were white, 40% indicated their political affiliation as independent, 27% as democrats, and 26% as republican. The average age of respondents was 45.6 with a median age of 46. Participants in this stage were sampled to match the U.S. distribution of both political affiliation and ethnicity. Participants were balanced across the two conditions: 256 participants were in the *CorrectFeed* condition, and 252 participants were in the *ErrorFeed* condition.

Again, our core tool for analysis was two multilevel regression models with post-level trust (level 1) clustered under participants (level 2), and we computed the same metrics of trust at the post and participant levels as we did in study 1.

Study 2 results

In Table 3, we show the results from two mixed-effects models with random intercepts: one for true headlines and one for false headlines. This analysis only used the data from the evaluation phase, in which all true headlines had warnings attached and none of the false headlines had warning attached. In this table, we show the results of the *CorrectFeed* condition in reference to the *ErrorFeed* condition. Again, following the model used in Study 1, the condition is the independent variable, and post-level trust is the dependent variable with repeated measures

grouped by participants. In Fig. 4, we show the box plots for post-level and participant-level trust across both conditions.

The results in Table 3a show that participants who went through the *CorrectFeed* habituation trusted true headlines with erroneously placed warning labels significantly ***less*** than those who went through the *ErrorFeed* habitation. In other words, those participants learned, during the habituation phase, to trust the warning labels sufficiently to reconsider their own information veracity judgement. They adopted the warning label heuristic. Although the magnitude of this effect was small, this result demonstrates that participants formed ecologically rational heuristics during the habituation phase and that those ERH spilled over into the evaluation phase, despite the change in condition. Taken together with the results of study 1, the results further show that information consumers can make ecologically rational adaptations when given sufficient exposure to the new media environment. The results in Table 3b show that there was no significant difference in participants' trust in false, unlabeled headlines across the different habitation phases. This result is similar to what we found in study 1, that the heuristic is more likely directly tied to the presence of the warning label rather than its absence. We thus continue to study 3, which includes placing labels (warning or supporting) on *all* news headlines.

Study 3: How sensitive are people to warning labels?

So far, we demonstrated that erroneously placed warning labels significantly shifted trust in information, both true and false, with somewhat more significant results for warning labels erroneously placed on true headlines. We argued that this broad-based shift was due to information consumers making ecologically rational adaptations to previously experienced media environments, where warning labels are the default heuristic in trust decisions. While adhering to warning labels for trust decisions is a desired result when those warnings are correct, a byproduct of this strategy is distrusting true information when it is labeled by mistake.

In study 3, we attempt to estimate the effect of erroneous labels by examining news consumers' sensitivity to warning labels, using signal detection analysis. Signal detection theory (SDT) applies to situations where two stimulus types exist -signal and noise- and should be discriminated[45]. The theory has been applied to situations similar to ours, for example to lie detection, information retrieval relevance, and factors in misinformation susceptibility[41,45]. In its simplest form, the theory is applied to yes/no response systems where in a series of trials, the signal is sometimes present in the noise and the respondent answers 'yes' for signal present and 'no' otherwise[46]. Signal detection analysis enables us to compute the respondents' sensitivity parameter (d'), which represents the extent to which the person can differentiate between the signal and noise[47]. d' is computed using the proportion of 'hits' (correctly detecting the signal when present) to 'false alarms' (incorrectly detecting the signal when it is absent)[46,47]. Higher values of d' mean better ability to detect the signal, or to separate the signal from noise.

In this study, we employed Signal Detection Analysis for a rating type system—rather than a binary yes/no system—where responses are provided on an ordinal scale[45]. We examine the ROC curve of each condition to understand respondents' ability to detect the signal. For this analysis, we defined the following:

- Signal present: false news headline
- Signal absent: true news headline
- Hit: correctly distrusting false news

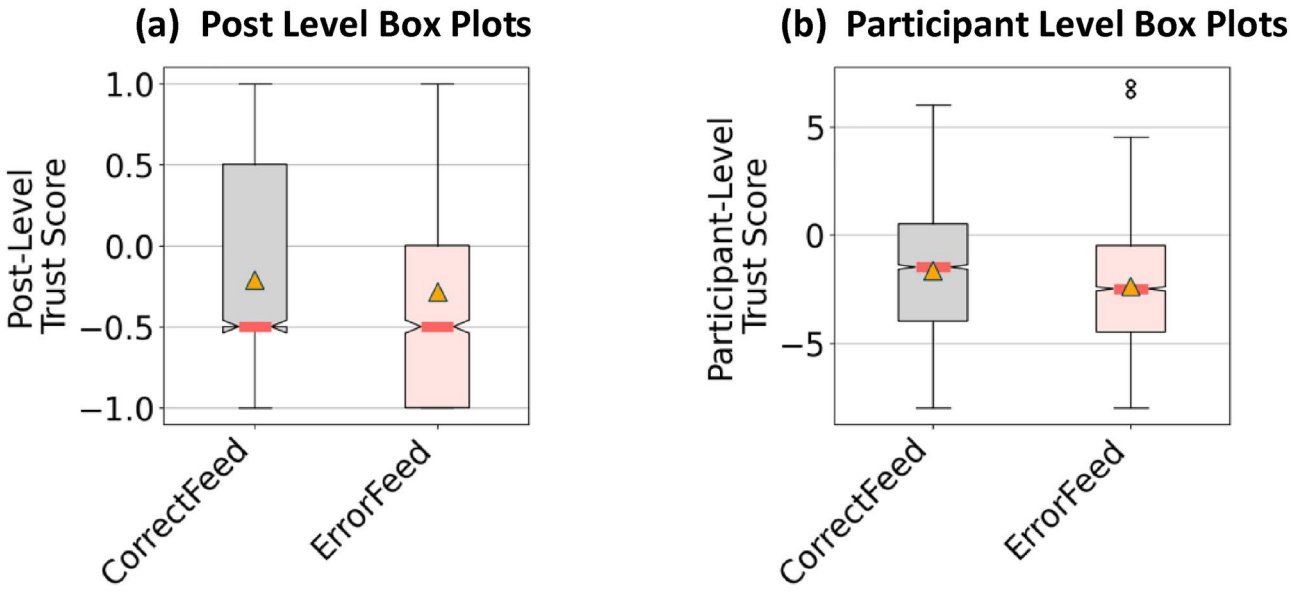

Fig. 4. Distributions of trust in true headlines during the evaluation stage at (**a**) post and (**b**) participant levels across the ErrorFeed and CorrectFeed conditions. Note, true headlines were labeled with warning labels during the evaluation stage.

- False alarm: incorrectly distrusting true news

We conducted a four-arm between-subjects experiment, in which participants (n = 429) again rated their trust in true and false headlines. While the third study mirrored study 1 in its basic task, there are several key differences. The third study had four conditions. There was a control condition—in which no labels were shown, there was a *random* label condition—in which labels were assigned to posts uniformly at random, there was an *error* label condition—in which all labels were incorrectly placed on headlines and finally, there was a *correct* label condition- in which all labels were correctly placed on headlines. Unlike in the prior two studies, in this study *all* posts were labeled. Specifically, we included not only a warning label of a post being false, but we also included a label stating that the post was true. An example of these labels can be found in Fig. 5. This design is notably different from other research in this area and from real-life moderation systems, which currently only label false information. We chose this design to test participants' sensitivity to labels across both veracity and label correctness.

In all conditions, each participant saw 20 posts (making 8,580 headline-level data points), which were randomly selected from a set of 30 posts – half of which were false. The 20 posts that were shown were balanced between true and false headlines. Headlines were again sampled as in the prior two studies, and they were resampled to ensure they were recent headlines.

The study was launched on Prolific and completed in early February 2025. In terms of demographics, about 50% identified as female and 50% as male, 56% of respondents were white, 38% identified as independents, 29% as Republicans, and 30% as Democrats. The average age of participants was 43.81 (median 44). Participants in this study were sampled to match the U.S. distribution of political affiliation and ethnicity. Participants were balanced across the four conditions: 111 participants were assigned to the control condition, 110 participants were assigned to the random label condition, 110 participants were assigned to the error label condition, and 108 were assigned to the correct label condition. Two new metrics are used in Study 3. First, since we employed a rating task, rather than a yes/no response system, we examined the ROC curve for each condition, which distinguishes hits from false alarms[45]. This method is akin to the use of ROC curves when comparing classifiers in Machine Learning experiments (e.g. Ref.[48]). In addition, we binned the ordinal scale into a binary yes/no scale, where trust is a combination of *somewhat trust and completely trust,* and distrust is a combination of *somewhat distrust and completely distrust*. This binning allowed us to compute the d' parameter for each condition using the following equation[46]:

$$d' = z(FA) - z(H),$$

where the FA is distrust in true news and H is distrusting in false news.

This d' metric is computed for each person in the experiment. Together, these two metrics allow us to clearly interpret the improvements and degradations from warning label systems of differing accuracy.

Study 3 results

Figure 6 shows the distribution of d' and the ROC curves for study 3. The average d' values for the four conditions in study 3 were:

- Correct Label 0.7603
- Control 0.4418
- Random Label 0.2513

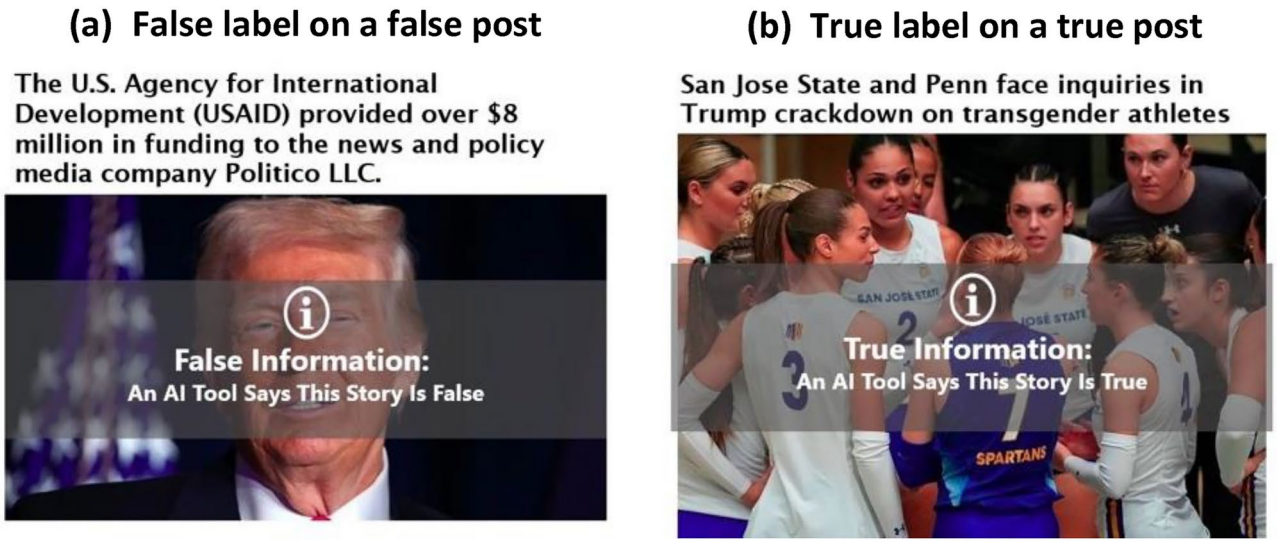

Fig. 5. Example stimuli from study 3, where both false and true labels were presented.

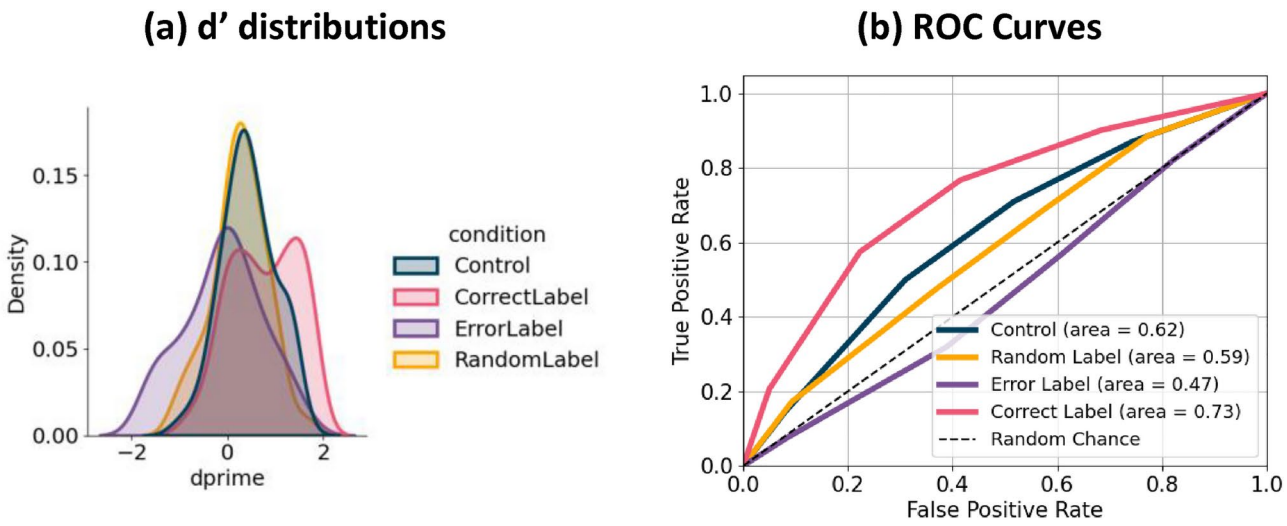

Fig. 6. (**a**) d' distributions across the four conditions of study 3. (**b**) ROC curves across the four conditions of study 3. Note that some studies which use SDT show the distributions of false alarms and hits. We are showing the distributions of d'—the difference between the false alarm rate and hit rate. The higher the d' score, the better ability to detect the signal, or to separate the signal from noise.

- Error Label -0.0855

The same order can also be seen in the area under the respective ROC curves (Fig. 6b). What these results show is that respondents were best able to detect the signal (i.e. distrust false news) when the labels were correctly placed, servicing to boost their own ability to detect false news. This result provides further support that participants relied on the warning labels as a useful heuristic. Further, in both the error label condition (where all labels were completely wrong) and the random label condition, respondents' ability to detect the signal was lower than the control condition, meaning that the erroneous labels reduced their ability to correctly detect the signal. In fact, in the error label condition, respondents' ability to detect the signal was worse than random chance.

In terms of magnitude, we can use the average d' values to compare the impact of the different labeling conditions. Specifically, the boosting (all correct) condition improved the average d' over the control condition by 0.3185 ($p < 0.001$), whereas the completely wrong labeling condition reduced the average d' from the control group by 0.5273 ($p < 0.001$), indicating a greater negative effect of wrong labels. Further, the random condition also caused a reduction in the d' value from the control condition ($p < 0.05$), indicating that the negative impact of wrong labels was greater than the positive impact of correct labels within the condition – despite those labels being approximately balanced between correct and incorrect placement.

Finally, we returned to the results of study 1 and employed signal detection analysis on these results as well (Fig. 7). As a reminder, study 1 had a similar design to study 3, except it only showed warning labels, not corroborating labels. The correct condition in study 1 placed warning labels on all false headlines and no labels on true headlines. The error condition placed warning labels on all true headlines and no labels on false headlines. The general order of the ROC curves for study 1 is similar to that of study 3, but we do see an improvement in the correct label condition from study 1 to study 3. This improvement can be attributed to placing the supporting labels on the true headlines, demonstrating an added value of labeling *all* news content and not just the false headlines.

Discussion

People develop ecologically rational heuristics leading them to trust the advice of news warning labels. If they develop these heuristics in one environment (e.g. human fact checking, one specific algorithm) but then apply it to a different environment (e.g. a less accurate algorithm), they may end up automatically trusting the advice *even if that advice is wrong.* Hence, misplaced content warning labels *can* significantly influence information trust decisions. Our results demonstrate that mislabeling information not only significantly reduces discernment ability from peoples' baseline discernment ability—when no labels are present -but can reduce discernment ability to worse than random chance.

Within the motivating scenario of using AI instead of humans to label content, this implication suggests that it is more important to thoughtfully consider what information needs to be labeled than it is to attempt to label all false information through automation. While our results do show that correct labels can significantly boost our ability to discern true and false information – particularly when true information is also labeled – perfectly labeling all information is likely impossible in real-life settings[17,18].

However, our results also suggest that people would adapt to content warning label systems that are not always correct. We expect that since information consumers can adapt to new media environments, the increase in erroneous decisions by those consumers after a switch to a more scalable and error-prone solution – like AI labeling—may only be a temporary phenomenon, and those erroneous decisions can be further mitigated by

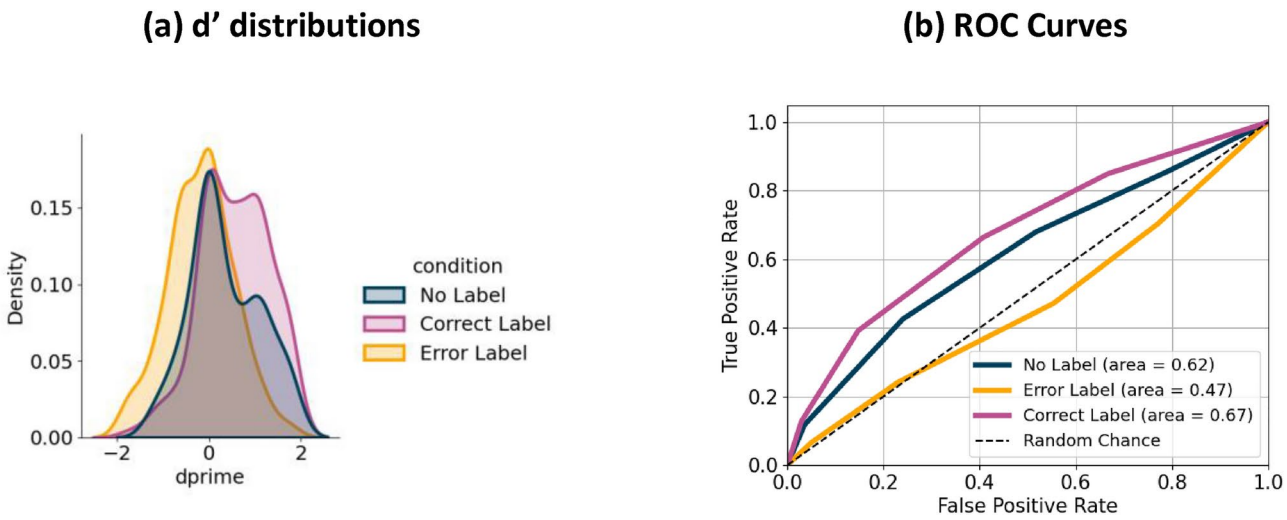

Fig. 7. (**a**) d' distributions across the three conditions of study 1. The average d' value for the *no label* condition was 0.3435, for the *correct label* condition was 0.5592, and for the *error label* condition was -0.1740. (**b**) ROC curves across the three conditions of study 1.

properly educating users on the algorithm's limitations. As consumers learn that the warning label system is not always correct, they may adjust their ecologically rational heuristics to the new system's accuracy, or they may rely on additional heuristics or systematic processing of the information. However, this would significantly weaken a powerful tool in our information intervention toolbox. Currently, warning labels are trusted, default decision heuristics with surprisingly broad effectiveness[3].

Based on the above, we caution that by optimizing our systems on scalability over correctness, we would reduce or even eliminate the usefulness of warning labels. We should also consider this result in light of the changing and emerging nature of news. The time lag during which prior ERH will still be employed to form a news judgement may be the critical time in which novel news stories are evaluated to form future opinions[39]. Therefore, it may be better to deploy a less scalable but highly effective intervention during these select, critical events – like a pandemic or election – than it is to deploy a less effective intervention across many events – both big and small. Given that people rarely engage in effortful information evaluation tasks and rarely process all aspects of information when making trust decisions[28], providing reliable short-cuts to critical information decisions should be seen as a top priority when developing interventions. A nuanced and tailored implementation of different moderation support systems – such as building tools for specific topics or events—may be an optimal strategy moving forward.

Our results also imply that within a single environment, uncommon errors can lead to incorrect decisions. Paradoxically, this result implies that the harm from a highly accurate system's *infrequent errors* may be as bad as the harm from a less accurate system with more errors – depending on the specific information that is mistakenly labeled. We term this the *rare failure effect*. The rare failure effect again suggests that interventions should be carefully and selectively deployed – whether those interventions are automated or not.

Future work should examine other factors that may interact with rational adaptations to media systems. For example, when it comes to algorithmic advice, two opposite reactions may arise when news consumers interact with the algorithms over time. Incorrect advice from automated systems can create algorithmic *aversion*. That is, people avoid algorithmic advice after they see the algorithm make a mistake[49]. In contrast, other studies have shown that incorrect advice from automated systems can negatively impact task performance across a variety of contexts due to an *overreliance* on algorithmic advice rather than an *aversion* to algorithmic advice[50]. In both cases – aversion and overreliance – one's experience with the task and one's experience with the system, whether that system is automated or not, play a role, as do many other factors, such as the decision's significance, the nature of the task being done, past performance of the algorithm, motivation and ability to process the information, and the level of human involvement in the outcome[51,52]. There are several important research questions stemming from this literature. For instance, does the perceived decision significance – trusting in information about an election versus information about a recent fad – moderate one's overreliance on warning labels? Or does the prior motivation or institutional trust of information consumers moderate the size of the *rare failure effect*? Further, participant-level features such as demographics and prior beliefs may interact with these effects. The experimental methods used in this paper can be extended to answer these questions.

Our results also demonstrated asymmetrical effects between erroneously labeling true and false information. That is, we find a larger effect in reducing information trust from warning labels erroneously placed on true information than warning labels correctly placed on false information. We suspect that at least part of this asymmetry may be due to floor effects. As shown in the control conditions for both study 1 and study 3, people trusted true information more and false information less by default. The ROC curves in Figs. 6 and 7 show that without warning labels people have some ability to discern between true and false information (AUC of 0.62 in both studies). Hence, trust in true information has more "room" to decrease than trust in false information

when a warning label is attached. However, there may be features of specific headlines that interact with one's trust in the headline, even if warning labels are ultimately followed. Given both the number of headlines used and the random selection of headlines used in our experiments, headline specific effects are averaged out. Still, prior work has provided some evidence that information features still play a role in trust even during warning label interventions[8]. Future work can examine the relationship between specific information and warning label errors with different experimental designs, such as adding information trait conditions to the experiment used in this paper.

The implications of our results extend well beyond the hypothetical setting of switching from human to AI fact checkers. For example, another direction for future research is to examine changes to other components in the news veracity environment. Content moderation within a single social media platform can change for many reasons, outside of implementing a new protocol or automation. For example, change in the ownership of a media platform can shift the priorities, affordances, and norms of the platform. When Twitter was bought by Elon Musk the validity of various credibility cues changed. For instance, the "blue check" that originally indicated that an account was verified to be a "notable" account holder—such as an organization or celebrity – became a 'paid for' feature available to anyone. The Trust and Safety Council—a board that provided guidance on content moderation—was dissolved, and content moderation was made significantly less strict[53]. Yet, many of the moderation tools, like Community Notes/BirdWatch, still exist in some form. Similarly, in January 2025, Meta announced that it will stop third-party fact checking on its platforms (Facebook and Instagram) in favor of crowd-based moderation like Community Notes. Our work demonstrates that users may be able to adapt to underlying changes in other credibility cues, not only direct warning labels (see Ref.[54] for another credibility cue), but that these adaptations come at a cost. It is worthwhile expanding the empirical studies to understand adaptation to different types of credibility cues.

Another context where these results may be relevant is in understanding the impacts of cross-platform use. Information consumers frequently use multiple social media platforms at the same time. Individual platforms may implement interventions differently, and those systems may change in design, correctness, and reliability over time. Our results suggest that information consumers' habits and heuristics developed on one platform may spill over to other platforms. If the warning label systems of both platforms are similar, this spillover may not be consequential. However, if the error rates are very different from each other, this spillover could be harmful. This implication supports the claim that platforms should collaborate on their moderation efforts[55]. Prior work has highlighted the importance of understanding the effects of both information flow and user migration across differing social media platforms. For example, we know that anti-social behavior and false information on one platform can spill over to other platforms, even if those platforms have different content moderation rules and norms[56,57]. Hence, the same type of effect may be true of information trust heuristics. There are many factors that likely interact with the migration of heuristics, such as the rate of use of one platform or another, the type of information consumed on one platform versus another, or even the information modality. Future work should design experiments to investigate the strength of spillover effects across platforms and the factors that may moderate or exacerbate those effects (Supplementary Information).

Conclusion

In this work we examined erroneously placed warning labels on news headlines. We show that people rely on warning labels in forming judgement on news veracity, even when those labels are incorrect. Further, we empirically show that this over-reliance on content labels is due to ecological rationality adaptations to our current media environments, where warning labels are human-generated and mostly correct. Our work contributes to the theory of ecologically rational heuristics by providing empirical evidence of the theory in a new context. Ecological rationality has been primarily studied in managerial information systems, such as predicting future job performance of employee candidates[30]. Our work illustrates that this theory applies to more than just predictions and decisions by managers but also to trust decisions by information consumers.

Data availability

The data supporting the findings of this study is available from the Open Science Framework: https://osf.io/wbkr3/?view_only=c775442cb13644fd8af6aae1ea92f4be.

Received: 3 October 2024; Accepted: 10 April 2025
Published online: 22 April 2025

References

1. Ecker, U. et al. Misinformation poses a bigger threat to democracy than you might think. *Nature* **630**, 29–32 (2024).
2. Morrow, G., Swire-Thompson, B., Polny, J. M., Kopec, M. & Wihbey, J. P. The emerging science of content labeling: Contextualizing social media content moderation. *J. Assoc. Inf. Sci. Technol.* **73**, 1365–1386 (2022).
3. Martel, C. & Rand, D. G. Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Curr. Opin. Psychol.* **54**, 101710 (2023).
4. Horne, B. D., Nevo, D., O'Donovan, J., Cho, J.-H. & Adalı, S. Rating reliability and bias in news articles: Does AI assistance help everyone?. *Proc. Int. AAAI Conf. Web Social Media* **13**, 247–256 (2019).
5. Epstein, Z. et al. Do explanations increase the effectiveness of AI-crowd generated fake news warnings?. *Proc. Int. AAAI Conf. Web Social Media* **16**, 183–193 (2022).
6. Clayton, K. et al. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Polit. Behav.* **42**, 1073–1095 (2020).
7. Porter, E. & Wood, T. J. Political misinformation and factual corrections on the Facebook news feed: Experimental evidence. *J. Polit.* **84**, 1812–1817 (2022).

8. Horne, B. D. Does the Source of a Warning Matter? Examining the Effectiveness of Veracity Warning Labels Across Warners. In *Proc. of the International AAAI Conference on Web and Social Media (Vol. 19)* (2025).
9. Meta. About fact-checking on Facebook, Instagram, and Threads. https://www.facebook.com/business/help/259358671757194 0?id=673052479947730 (2024).
10. Frangoul, A. With over 1 billion users, here's how YouTube is keeping pace with change. *Blogi. Päivitetty* **14**, 2018 (2018).
11. Zhou, X. & Zafarani, R. A survey of fake news: Fundamental theories, detection methods, and opportunities. *ACM Comput. Surv. (CSUR)* **53**, 1–40 (2020).
12. Shu, K., Wang, S. & Liu, H. Beyond news contents: The role of social context for fake news detection. in *Proc. of the twelfth ACM International Conference on Web Search and Data Mining* , 312–320 (2019).
13. Gruppi, M., Horne, B. D. & Adalı, S. Tell me who your friends are: Using content sharing behavior for news source veracity detection. Preprint at http://arXiv.org/arXiv:2101.10973 (2021).
14. Raza, S. & Ding, C. Fake news detection based on news content and social contexts: A transformer-based approach. *Int. J. Data Sci. Anal.* **13**, 335–362 (2022).
15. Yaqub, W., Kakhidze, O., Brockman, M. L., Memon, N. & Patil, S. Effects of credibility indicators on social media news sharing intent. In *Proc. of the 2020 Chi Conference on Human Factors in Computing Systems*, 1–14 (2020).
16. Wang, S. Factors related to user perceptions of artificial intelligence (AI)-based content moderation on social media. *Comput. Hum. Behav.* **149**, 107971 (2023).
17. Bozarth, L., Saraf, A. & Budak, C. Higher ground? How groundtruth labeling impacts our understanding of fake news about the 2016 US presidential nominees. *Proc. Int. AAAI Conf. Web Social Media* **14**, 48–59 (2020).
18. Horne, B. D., Nevo, D. & Smith, S. L. Ethical and safety considerations in automated fake news detection. *Behav. Inf. Technol.* https://doi.org/10.1080/0144929X.2023.2285949 (2023).
19. Freeze, M. et al. Fake claims of fake news: Political misinformation, warnings, and the tainted truth effect. *Polit. Behav.* **43**, 1433–1465 (2021).
20. Echterhoff, G., Groll, S. & Hirst, W. Tainted truth: Overcorrection for misinformation influence on eyewitness memory. *Soc. Cogn.* **25**, 367–409 (2007).
21. Pennycook, G., Bear, A., Collins, E. T. & Rand, D. G. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Manag. Sci.* **66**, 4944–4957 (2020).
22. Chaiken, S. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *J. Pers. Soc. Psychol.* **39**, 752 (1980).
23. Chaiken, S. & Ledgerwood, A. A theory of heuristic and systematic information processing. *Handb. Theor. Soc. Psychol.* **1**, 246–266 (2012).
24. Todorov, A., Chaiken, S. & Henderson, M. D. The heuristic-systematic model of social information processing. *Persuasion Handb. Dev. Theory Pract.* **23**, 195–211 (2002).
25. Chen, S., Duckworth, K. & Chaiken, S. Motivated heuristic and systematic processing. *Psychol. Inq.* **10**, 44–49 (1999).
26. Nickerson, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* **2**, 175–220 (1998).
27. Pennycook, G. & Rand, D. G. The psychology of fake news. *Trends Cogn. Sci.* **25**, 388–402 (2021).
28. Metzger, M. J. & Flanagin, A. J. Credibility and trust of information in online environments: The use of cognitive heuristics. *J. Pragmat.* **59**, 210–220 (2013).
29. Todd, P. M. & Gigerenzer, G. Environments that make us smart: Ecological rationality. *Curr. Dir. Psychol. Sci.* **16**, 167–171 (2007).
30. Luan, S., Reb, J. & Gigerenzer, G. Ecological rationality: Fast-and-frugal heuristics for managerial decision making under uncertainty. *Acad. Manag. J.* **62**, 1735–1759 (2019).
31. Zannettou, S. ' I Won the Election!': An empirical analysis of soft moderation interventions on Twitter. *Proc. Int. AAAI Conf. Web Social Media* **15**, 865–876 (2021).
32. Barrón-Cedeno, A., Jaradat, I., Da San Martino, G. & Nakov, P. Proppy: Organizing the news based on their propagandistic content. *Inf. Process. Manag.* **56**, 1849–1864 (2019).
33. Gruppi, M., Smeros, P., Adalı, S., Castillo, C. & Aberer, K. SciLander: Mapping the scientific news landscape. *Proc. Int. AAAI Conf. Web Social Media* **17**, 269–280 (2023).
34. Ciampaglia, G. L. et al. Computational fact checking from knowledge networks. *PLoS One* **10**, e0128193 (2015).
35. Yang, K.-C. & Menczer, F. Accuracy and political bias of news source credibility ratings by large language models. Preprint at https://arXiv.org/2304.00228 (2023).
36. Zhou, X., Sharma, A., Zhang, A. X. & Althoff, T. Correcting misinformation on social media with a large language model. Preprint at https://arXiv.org/2403.11169 (2024).
37. Reis, J. C. S., Correia, A., Murai, F., Veloso, A. & Benevenuto, F. Supervised learning for fake news detection. *IEEE Intell. Syst.* **34**, 76–81 (2019).
38. Douglas, B. D., Ewell, P. J. & Brauer, M. Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS One* **18**, e0279720 (2023).
39. Nevo, D. & Horne, B. D. How topic novelty impacts the effectiveness of news veracity interventions. *Commun. ACM* **65**, 68–75 (2022).
40. Martel, C. & Rand, D. G. Fact-checker warning labels are effective even for those who distrust fact-checkers. *Nat. Hum. Behav.* **8**, 1957–1967 (2024).
41. Sultan, M. et al. Susceptibility to online misinformation: A systematic meta-analysis of demographic and psychological factors. *Proc. Natl. Acad. Sci.* **121**, e2409329121 (2024).
42. Pennycook, G. & Rand, D. G. Accuracy prompts are a replicable and generalizable approach for reducing the spread of misinformation. *Nat. Commun.* **13**, 2333 (2022).
43. Epstein, Z., Sirlin, N., Arechar, A., Pennycook, G. & Rand, D. The social media context interferes with truth discernment. *Sci Adv* **9**, eabo6169 (2023).
44. Orchinik, R., Martel, C., Rand, D. G. & Bhui, R. Uncommon errors: adaptive intuitions in high-quality media environments increase susceptibility to misinformation. *PsyArXiv* 10, (2023).
45. Stanislaw, H. & Todorov, N. Calculation of signal detection theory measures. *Behav. Res. Methods Instrum. Comput.* **31**, 137–149 (1999).
46. Macmillan, N. A. Signal detection theory. *Stevens' Handb. Exp. Psychol. Methodol. Exp. Psychol.* **3**, 43–90 (2002).
47. Sorkin, R. D. & Woods, D. D. Systems with human monitors: A signal detection analysis. *Hum. Comput. Interact.* **1**, 49–75 (1985).
48. Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.* **30**, 1145–1159 (1997).
49. Dietvorst, B. J., Simmons, J. P. & Massey, C. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* **144**, 114 (2015).
50. Cecil, J., Lermer, E., Hudecek, M. F. C., Sauer, J. & Gaube, S. Explainability does not mitigate the negative impact of incorrect AI advice in a personnel selection task. *Sci. Rep.* **14**, 9736 (2024).
51. Dietvorst, B. J., Simmons, J. P. & Massey, C. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Manag. Sci.* **64**, 1155–1170 (2018).
52. Molina, M. D. & Sundar, S. S. Does distrust in humans predict greater trust in AI? Role of individual differences in user responses to content moderation. *New Media Soc.* **26**, 3638–3656 (2024).

53. Hickey, D. et al. Auditing Elon Musk's impact on hate speech and bots. *Proc. Int. AAAI Conf. Web Social Media* **17**, 1133–1137 (2023).
54. Celadin, T., Capraro, V., Pennycook, G. & Rand, D. G. Displaying news source trustworthiness ratings reduces sharing intentions for false news posts. *J. Online Trust Saf.* https://doi.org/10.54501/jots.v1i5.100 (2023).
55. Wilson, T. & Starbird, K. Cross-platform disinformation campaigns: Lessons learned and next steps. *Harvard Kennedy School Misinf. Rev.* https://doi.org/10.37016/mr-2020-002 (2020).
56. Childs, M., Buntain, C., Z. Trujillo, M. & D. Horne, B. Characterizing youtube and bitchute content and mobilizers during us election fraud discussions on twitter. In *Proc. of the 14th ACM Web Science Conference 2022*, 250–259 (2022).
57. Russo, G., Verginer, L., Ribeiro, M. H. & Casiraghi, G. Spillover of antisocial behavior from fringe platforms: The unintended consequences of community banning. *Proc. Int. AAAI Conf. Web Social Media* **17**, 742–753 (2023).
58. Greene, C. M. et al. Best practices for ethical conduct of misinformation Research. *Eur. Psychol.* https://doi.org/10.1027/1016-9040/a000491 (2022).

Author contributions

B.H.: conceptualization, methodology, data curation, formal analysis, writing-original draft, writing-review and editing. D.N.: conceptualization, methodology, data curation, writing-original draft, writing-review and editing.

Declarations

Competing interests

The authors declare no competing interests.

Ethical approval, informed consent and debriefing

All research was performed in accordance with relevant guidelines/regulations applicable when human participants are involved (Declaration of Helsinki). The research was approved by Rensselaer Polytechnic Institute Human Research Protections Program (HRPP), which determined that the application was eligible for exempt status under 45 CFR 46.104.d, Category 2 (https://www.hhs.gov/ohrp/regulations-and-policy/regulations/45-cfr-46/revised-common-rule-regulatory-text/index.html#46.101). Written informed consent was obtained from all subjects. All subjects were 18 years old or above. We followed best practices for conducting misinformation research[58]. For instance, given the experiment used real information, all participants went through a debriefing, in which they were presented with information about the experiment, the veracity of headlines, and an explanation of how veracity was determined for each headline.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-98221-7.

Correspondence and requests for materials should be addressed to B.D.H.