Proceedings of the Fifteenth International AAAI Conference on Web and Social Media (ICWSM 2021)

Machine Learning Explanations to Prevent Overtrust in Fake News Detection

Sina Mohseni[1], Fan Yang[1], Shiva Pentyala[1], Mengnan Du[1], Yi Liu[1], Nic Lupfer[1], Xia Hu[1], Shuiwang Ji[1], Eric Ragan[2]
[1] Texas A&M University, College Station, TX
[2] University of Florida, Gainesville, FL
{sina.mohseni,nacoyang,pk123,dumengnan,yiliu,nlupfer,xiahu,sji}@tamu.edu, eragan@ufl.edu

Abstract

Combating fake news and misinformation propagation is a challenging task in the post-truth era. News feed and search algorithms could potentially lead to unintentional large-scale propagation of false and fabricated information with users being exposed to algorithmically selected false content. Our research investigates the effects of an *Explainable AI* assistant embedded in news review platforms for combating the propagation of fake news. We design a news reviewing and sharing interface, create a dataset of news stories, and train four interpretable fake news detection algorithms to study the effects of algorithmic transparency on end-users. We present evaluation results and analysis from multiple controlled crowdsourced studies. For a deeper understanding of *Explainable AI* systems, we discuss interactions between user engagement, mental model, trust, and performance measures in the process of explaining. The study results indicate that explanations helped participants to build appropriate mental models of the intelligent assistants in different conditions and adjust their trust according to their perceptions of model limitations.

Introduction

Artificial intelligence (AI) algorithms are used in a variety of online applications from product recommendation and targeted advertisement to loan and insurance rate prediction. However, as AI-based decision-making is directly affecting people's lives, the accountability and fairness of advanced AI algorithms are under question. In recent years, the need for algorithmic transparency is gaining more attention to enable accountable AI-based decision-making systems. To this end, *Explainable AI* (XAI) techniques (Gunning 2017) have been introduced to annex transparency into black-box machine-learning algorithms. Interpretability can help users to build a mental model of how algorithms work and build appropriate trust in intelligent systems Rader et al. (2018).

In the social media domain, news feed and search algorithms function similar to decision-making algorithms, as users are exposed to algorithmically selected content (Trielli and Diakopoulos 2019). Blindly trusting algorithmically-curated news could potentially lead to unintentional large-scale propagation of false and fabricated information with users being exposed to false content and its re-sharing through social media. Human review of news and data mining techniques for fake-news detection and debunking are commonly being practiced as primary approaches to reduce fake news in social media. However, reviewing the life cycle of news in social media reveals opportunities to combat the propagation of fake news within news-feed platforms (Mohseni, Ragan, and Hu 2019). For example, AI-based news review assistant tools can be embedded in news feed platforms and have the potential to benefit users by providing direct suggestions related to news credibility rather than automatic organizational news debunking. Nevertheless, proposed AI-based news assistants could raise concerns about possible algorithmic biases and errors causing distrust for users. Therefore, machine learning explanations could be used as tools to help users to build justified trust in intelligent assistants.

Our research objective is to study the effects of transparency for fake news detection through an XAI assistant for news review applications and social media. We investigate whether the interpretability of the fake news detector algorithm could enhance users' overall experience and result in increased credibility of user-shared news. We also aim to examine whether model explanations can help users to avoid overtrusting the fake news detector when explanations are nonsensical to users. We formulate our goals into the following research questions:

- RQ1: Do AI and XAI assistants help end-users share more credible news?
- RQ2: How do AI explanations affect users' mental models of intelligent assistants?
- RQ3: How do AI explanations affect end-users' trust and reliance in intelligent assistants?

To address these research questions, we designed a news reviewing and sharing interface with a built-in interpretable fake news detector (AI and XAI assistants) for end-users and run a series of evaluation experiments. The study results indicate the complexity of the fake news detection problem and the limitations of current model interpretability techniques for this task. Though the addition of explanations to our system did not improve user task performance, we ob-

served that explanations helped participants' to build appropriate mental models of the intelligent assistants in different conditions and adjust their trust accordingly based on their perception of model logic.

Related Work

Machine learning algorithms are heavily used in online platforms and social media to analyze user data for improving user experience and increasing corporate profit. However, the lack of transparency can raise data privacy and model trustworthiness concerns in critical domains, and hence potentially decreases user trust and confidence in the long run (Mohseni, Zarei, and Ragan 2019). In this regard, researchers study the communication of algorithmic processes in various domains such as online advertising (Eslami et al. 2018), social media feeds (Eslami et al. 2015), and personalized news search engines (ter Hoeve et al. 2017). This section briefly reviews machine learning and human-computer interaction papers related to the explainable news feed and fake news detection systems.

Interpretable Fake News Detection

Reviewing machine learning techniques for misinformation detection shows diverse directions like style-based approaches to detect deception-oriented writing (Rubin and Lukoianova 2015) and hyper-partisan content (Potthast et al. 2017). To incorporate more data for representation learning, Shu et al. (2017) included news publisher information, user stance, and user engagement together in their Tri-Relationship fake news detection framework. In other work, Popat et al. (2018) used the Google search engine to directly collect similar instances from the web with a sentence similarity as the measure. Their technique leverages external news articles to gather evidence from online sources.

Multiple research efforts have shown that algorithm interpretability could potentially improve user attitudes toward algorithms. Du et al. (2019) categorized interpretation methods to explain predictions of Natural Language Processing (NLP) models into four categories of back-propagation based, perturbation based, local approximation based, and decomposition-based techniques. Back-propagation based methods calculate gradients or variants of gradients of a model prediction with respect to each input (Hechtlinger 2016). Gradient values for each word in the input can represent its contribution to the model prediction. Alternatively, in perturbation based techniques, the occlusion of input text can cause changes in the model prediction resulting to estimate word contributions (Li, Monroe, and Jurafsky 2016). Unlike the other two techniques, decomposition-based methods can model the data flow process by calculating the additive contribution of each input word to final prediction (Murdoch, Liu, and Yu 2018). Lastly, local approximation-based methods can explain a complex model's predictions by approximating its behavior around an input instance (Ribeiro, Singh, and Guestrin 2016).

Recent interpretable fake news systems have shown advantages of interpretability for fake news detection, including helping end-users to find model weaknesses so that users can build an appropriate level of trust in model predictions. For instance, XFake detector in (Yang et al. 2019) uses a tree-based model visualization to explain the overall decision paths for input news instances. In another paper, Shu et al. (2019) provide feature-based explanations for important user comments from relevant news articles for user interpretability on fake news detection. However, While these models achieve moderate performance (i.e., below 80% detection accuracy) in detecting fake news, it remains uncertain that how would model explanations help end-users in detecting fake news. In this paper, we designed a news review environment with built-in explainable to study the effects of algorithmic transparency on human-AI collaboration for fake news detection.

Explainability for Users

Multiple studies on transparent AI explore design choices to build accurate mental model of algorithms and adjust end-users trust in AI systems. For instance, Kocielnik et al. (2019) investigate the accuracy indicator, example-based explanation, and user control as design choices to improve human-AI collaboration. Their findings indicate that users' perception of control had a significant positive effect on user trust. In the evaluation of XAI interfaces, Poursabzi et al. (2018) present a comprehensive evaluation study for users' mental models (via user prediction task) and trust (via user agreement with AI) in interpretable models. Their results indicate the positive effect of interpretability on participants' mental models, however, they did not observe improvement on user trust. On the other hand, Papenmeier et al. (2019) show users could potentially lose trust in AI when exposed to low fidelity explanations.

To gain insight into whether news recommendation algorithms should be transparent about their decisions, Hoeve et al. (2017) run a survey and learn that a large majority of respondents prefer explanations. However, in a follow up A/B testing, they find participants are not opening (via click count) model explanations. This could be due to the low urgency of explanations in news recommendation and/or their study news test set. In human studies for AI-based news fact-checking, Horne et al. (2019) run an experimental human subject study and find that feature-based explanations in AI assistant significantly improve users perception of news bias. Their measured effect size was much larger for participants who were frequent newsreaders and those familiar with politics, though. In another paper, Nguyen et al. (2018) present the design and evaluation of a mixed-initiative fact-checking system to blend human knowledge with machine learning algorithms. They also conclude that transparency and interactivity significantly affect users' ability to predict the veracity of given claims. To continue this line of research, we investigate how different types of model explanations affect the credibility of news shared by users in a social media like scenario. We also measure a wider range of explicit and implicit user feedback to study interactions among key XAI design goals in the explaining process.

System Design

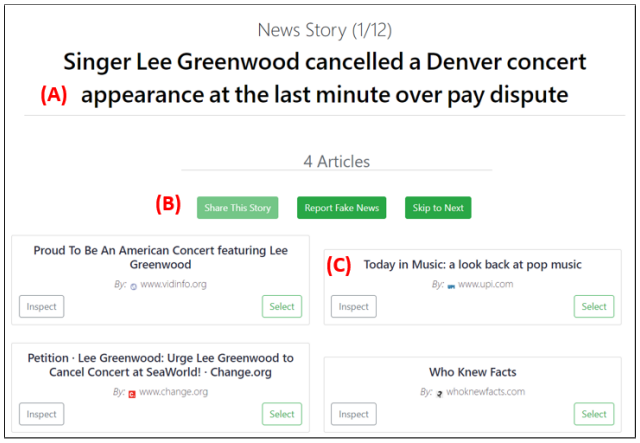
To serve our research goals for studying the role of interpretable models in fake news detection, we designed an interface for users to review news stories and articles as well as interact with a fake news detection assistant and its explanations. Our system's targeted users are the general public who read daily news and are not AI experts nor news analysts. We started with identifying candidates for useful and impactful explanations for fake news detection such as keyword attention, supporting evidence, and source credibility based on machine learning research on misinformation detection and human-computer interaction research on news feed systems (Mohseni, Ragan, and Hu 2019). We followed Mohseni et al.'s (2019) design framework for XAI systems, which involved design iterations with series of pilot user testing for both model explanations and news interface.

Explainable Interface

We designed an interface for users to review a queue of news stories, share true news for other users, and report fake news stories. Our interface design process started with multiple interface sketches that suit the news reviewing task. We aimed to design a simple interface with useful explanations for fake news detection. We tested mock-up implementations from the top design choices with a small number of participants. After reviewing feedback from user observations and interviews, we selected the most comprehensible and conclusive design for our human-subject experiments.

Figure 1-Top shows our baseline interface that enables the participants' news review task. The interface shows a news headline for a news story on the top (Figure 1-A) followed by a list of related articles below (Figure 1-C). The related articles provide context and article sources for the news headline, and they can help the user to understand contributing information and factors for model prediction. The system allows users to open and read the related articles, but for our study, it was not required for sharing the news headline. The system was designed to allow users to review news stories one-by-one and decide if 1) the story is true to be shared with other users, or 2) it is fake news to be reported, or 3) they want to skip to the next story due to their unfamiliarity with the topic or lack of confidence (see Figure 1-C).

Figure 1-Bottom shows our interface with the XAI assistant. The fake news detection assistant is embedded in our interface which provides the model prediction (with or without explanation) about the news story's credibility. Both the AI assistant prediction and its explanations are in the form of on-demand recommendations for the user, which are collapsible on user click. Our XAI assistant design provides *why-type* explanations (see (Mohseni, Zarei, and Ragan 2019) for discussion of explanation types) from the ensemble of fake news detectors. These explanations describe the attribution of different news features (i.e., news headline, article text, and article source) for each news veracity prediction. These different explanations are presented to the user with the following visual elements: (1) A heatmap of keyword attribution score that explains how the XAI assistant learned word-level features in the news headline (Figure 1-D) and its related articles (in the news article page).



Interface in Baseline condition.

Interface in XAI-all condition.

Figure 1: Our news review interface with AI and XAI assistants. Interface components are removed for different study conditions. **Top:** Baseline interface without AI assistant. (A) news headline. (B) user selecting to share, report, or skip the news story. (C) a list of related news articles for the headline. The user can open articles to inspect details. **Bottom:** Interface with clickable model prediction and different feature attribution explanations. (D) a heatmap of word level feature attribution explanation for news headline. User can see the attribution score values in tooltips when hovering mouse over the keywords. (E) fake news prediction and confidence. (F) confidence for the headline and article separately. (G) a bar chart for each article attribution score in comparison to other related articles. Bar charts show lower values when the articles are less related to the headline or less significant for model prediction. (H) A donut chart for each news article for source attribution scores compared to headline (Claim) and article (Text) content.

(2) A single bar chart for each related article (Figure 1-G) explaining each related article's attribution score in comparison to other articles for the model prediction. (3) A pie chart to present attribution score for the articles' source in comparison to the articles' content attribution and news headline attribution. (4) A list of top-3 important sentences for the article is shown when reviewing news articles to explain sentence-level feature learning of the models.

The core design rationale for the four explanations was to embed model attribution explanations using visual elements for each news feature. Each visual explanation element was tested with users and refined through iterative design.

Fake News Data

Training data for our models come from two sources: a) news story headlines and labels from Snopes (www.snopes.com) and b) related articles crawled from Google search results (top 16). The related articles were collected for each Snope news headline separately and labeled the same as their respective Snopes news story statements with noisy label assumption for the purpose of model training. The training data includes 4638 news story headlines with an average length of 15 words and 30599 related articles with an average length of 1012 words. The dataset is balanced with half news stories labeled as True and the other half labeled as fake. We used 80% of data for model training, 10% data for validation, and 10% data for testing. We took news samples and model predictions from our test set to feed the interface for human-subject studies. Our dataset consists of news, rumors, and hoax which covers a range of different topics, including politics (725 stories), business (224 stories), health (192 stories), and crime (141 stories).

Interpretable Models

Following the previous implementations of NLP algorithms for fake news detection, we implement an ensemble of four classifiers for fake news detection to generate different types of explanations. Our purpose in choosing the ensemble model approach was to study the effects of different explanation types later in the evaluation experiments. First, the four models are trained separately in a standard training setup with cross-validation. Next, the ensemble of models is tuned at once using the validation set to optimize each model's weight for the overall performance. We obtain the final prediction score through averaged ensemble results with 73.65% detection accuracy in the test set. We review the details of each model in the following.

Our first model is a Bi-LSTM network (Huang, Xu, and Yu 2015) with an additional self-attention layer to extract attention scores for instance explanations. This model is trained on news headlines only and generates attention explanations for its predictions. In our empirical tests, Bi-LSTM network outperformed similar networks (e.g., RNN, LSTM. RCNN) for our dataset by capturing both forward and backward states. We also use Word2Vec (Mikolov et al. 2013) to embed each word into an embedding vector before feeding into the network. This trained model achieves 72.00% fake news detection accuracy on our test set.

The second model performs fake news detection based on both the news story headlines and the set of related articles for each. The article set representation is constructed using hierarchical attention at sentence level and article level. We use the hierarchical attention network (HAN) (Yang et al. 2016) to help our model focus on the salient sentences and articles at two levels. HAN scores each article and selects the most important sentences in each article. Each sentence representation of an input article is generated by taking an average of the word embedding of all the words therein. Our design allows us to get the attribution score for each article and select the three most important sentences in each article using attention weighs. For the news story representation, similar to our first model, we used a Bi-LSTM network. Finally, a weighted sum is performed over all articles to build the article representation, which is combined with the news story representation to form the final vector representation for news story classification. This model achieved 76.04% classification accuracy on the test set.

For the third model, we use a knowledge distillation approach (Hinton, Vinyals, and Dean 2015) to approximate a deep architecture (teacher) with a random forest (student) model. This model takes news stories, related articles, and article source as the input, and with the mimic learning framework, we can leverage the performance of a deep model and analyze the attribute importance of news stories, articles, and their sources for each prediction. We first train a Bi-LSTM teacher model using Glove word embedding (Pennington, Socher, and Manning 2014) and then train a 60 trees XGBoost (Chen and Guestrin 2016) student model. The XGBoost student model provides attribute importance (news story headline, article content, and article source) as a form of instance explanations. Our third model achieved *72.08%* prediction accuracy in the test set.

For the last model, we use both news headlines and related articles to train a Bi-LSTM network with Word2Vec word embedding. We use an attention mechanism to focus on parts of the articles that are more relevant to the news story. In order to do so, we calculate a weighted average of the hidden state representation based on the attention score corresponding to all the article tokens (Kumar and Rastogi 2019). Our method then aggregates all the information about the news story, article context, and attention weights to predict the story's credibility. Finally, to generate an overall credibility label for the classification task, the final representation is processed using the final fully connected layer. The attention mechanism also generates keyword attribution explanations for each article.

Experimental Design

We design controlled human subject studies in order to test our hypothesis regarding the effectiveness of AI assistance and its explanation in the news review task. We present our study design details in terms of study conditions, evaluation measures, and participants' task.

Study Design

We conducted human-subjects studies for controlled comparison of elements of the AI assistant and its explanations.

Study Condition	Model Output	Model Explanations
Baseline	–	–
AI Assistant	Prediction and Confidence	–
XAI Assistants (3 conditions)	Prediction and Confidence	XAI-attention: Keywork importance heatmap for news headline and articles.
		XAI-attribution: News attribute and article importance for related articles.
		XAI-all: Explanations from both XAI-attribute and XAI-attention conditions.

Table 1: Study conditions and intelligent assistant components to detect fake news and explain its prediction.

The study followed a between-subjects design with five different conditions, where each participant used one variation of the news reviewing system as described in the following and summarized in Table **??**.

Baseline Condition: For the *Baseline* condition, we remove AI prediction and its explanations in the interface. The baseline interface (Figure 1, top) allows participants to review and share news headlines without any machine learning support. This condition serves as the baseline for human-alone performance in comparison to human-AI collaboration. Also, since the *Baseline* condition did not include AI or XAI elements, this condition did not measure the participants' mental model and trust in AI or XAI.

AI Assistant: Our interface in *AI Assistant* condition includes AI prediction and confidence for news headline credibility. Note that this condition used the same ensemble model as the XAI Assistant conditions but without showing the explanations. The AI predictions are in the form of on-demand using a collapsible menu on user click. This condition serves as the baseline for participants' mental model and trust measurements in the AI without explanation. Figure 1 shows model prediction and confidence at (E) and models confidence for the headline and articles separately at (F).

XAI Assistants: The user interface in *XAI assistant* conditions provides instance explanations in addition to news credibility prediction. We design three *XAI Assistant* conditions to study how different types of explanations affect Human-AI collaboration. We use two interpretable models in each *XAI Assistant* condition. The *XAI-attention* condition presents a heatmap of keywords using attention weights for news headline (Figure 1-D) and each related news articles. The *XAI-attribution* condition shows news attribution explanations for related articles and news sources. The hierarchical attention network generates articles importance score (Figure 1-G) and top-3 important sentences from each article. Our mimic model generates source, article, and news story attribution score (Figure 1-H) to present instance explanations. The *XAI-all* condition is the combination of explanations in the *XAI-attribution* and *XAI-attention* conditions. The purpose of designing *XAI-all* condition was to study the effect of variety of explanation types on participants.

Study Measures

We take users' mental model, human-AI performance, and trust as the primary measures in our studies. We mainly use quantitative methods for our measurements to aim for investigating the initial research questions (RQ1 – RQ3).

Task Performance: We calculate the veracity of participants' final shared and reported news as the main performance metric. We take the credibility score of user shared news as the number of shared true news divided by total shared news (equal to 12 in all experiments). We also review and analyze results for the incredibility score (calculated as 1.0 – *credibility score*) of all reported fake news as the secondary performance measures.

Mental Model: We take participants' accuracy in guessing model output (similar to Poursabzi et al. (2018)) as representative for model predictability and user mental model. For the measurement of this prediction task, we use short pop-up questions during the study to ask "what would the AI fake news detector predict for this news story?" from participants. Participants could respond with short "True" or "Fake" answers. Since we expect participants to interact and understand the intelligent assistant during the early stages of the study, the pop-up questions for mental model measurements were limited to the final third of the study (i.e., the last four news review instances).

User Trust and Reliance: We measure user trust using a subjective rating of participants' perceived accuracy of AI assistant. Specifically, participants answer "What was the accuracy of the AI fake news detection?" using a continuous slider bar (between 0–100%) to indicate their perception of AI or XAI assistant's accuracy in the post-study survey.

We also measure user reliance using participants' agreement rate with AI assistant predictions. To quantitatively measure participants' reliance on model predictions, similar to (Yin, Wortman Vaughan, and Wallach 2019), we calculate user agreement rate as the number of news stories which the participant inspected and agreed with the model prediction (either true or fake news), divided by the total number of shared or reported news stories.

Study Procedure

Participants started the task by accepting the information sheet including our approved IRB number and study contact points information. Next, participants saw step-by-step task

instructions with visual guides for all interface components. Visual instructions include descriptions for the headline and article attribution explanations from XAI assistant. Next, participants answered the pre-study questionnaire including text entry and multiple-choice questions. Participants then started the main task by reviewing news stories.

User Task: Participants were prompted to review a queue of news stories and share 12 true news for social media users. To engage participants to review news articles and their explanations, users had to select at least one article that represents the news headline for each news story they chose to share. They could always skip to the next news story (as many times as needed) if they were not familiar with the topic. The choice of the sharing task and ability to skip unfamiliar topics (unlike work that assumes participants are familiar with a short curated list of news stories e.g., (Horne et al. 2019; Nguyen et al. 2018)) improves the fake news detection task by allowing participants to interact and examine the AI/XAI assistant rather than focusing on news analysis. Participants also had the chance to flag news stories as fake if they found headlines to be fake; however, these were not counted toward the required number of shared stories needed for task completion. Also, in contrast with previous work, our interface gives a list of related news articles to provide the context of news stories for users. Further, unlike (Nguyen et al. 2018), participants did not receive feedback of the ground truth after each instance (i.e., whether the model made a correct or wrong prediction) to simulate a real-world scenario in which users do not have immediate access to the credibility of their daily news. During the last four news stories (the last third of the study), participants were asked pop-up questions about the AI assistant's prediction before revealing the model prediction; This was done to collect data to estimate user ability to predict the AI's output.

In the end, participants answered a final questionnaire of Likert-scale and slider questions about the AI assistant followed by four open-ended response forms.

Participant Pool

Our XAI system and study were designed for non-expert end-users with little knowledge of AI. We recruited remote participants from Amazon Mechanical Turk "Master" users with above 90% acceptance rate. To encourage participants to spend enough time on the task, we measured task duration and paid flexible time-based compensations. The payment was set to $10 per hour, and each participant could only participate once in the task. To further ensure data quality for analysis, we filtered data samples based on collected user engagement measures including task duration, number of clicks, and character counts in the final questionnaire form.

Experiments and Results

We ran five between-subject experiments in different interface conditions for hypothesis testing. The study had a total of 220 Amazon Mechanical Turk participants, which were distributed equally across conditions. From participants' self-reports, 47% were female, with 37.8% between 30-39 years old, 30.3% between 40-49 years old, 15.9% between 20-29 years old, and 3% between 50-59 years old. In terms of education, 51% had a bachelor's degree, 23% had a college degree less than bachelor, 14% had graduate school degree, 14% had high school education, and 1% had less than high school education. We removed data from 19 participants who spent less than 10 minutes or had especially low interaction behavior during the task. A total of 122.8 hours of study time was recorded for the remaining 201 participants, who on average spent 32.1 minutes (range = [10.3, 90.6] with SD = 20.3) on news review and selection, and 6.5 minutes answering surveys and reading instructions.

In all experiments, we used the same pre-processed and curated queue of news stories to eliminate potential confounds of different news inputs on experiment groups. The news stories' composition was organized to show one fake news for every true news to present a scenario with an equal mix of true and fake news. Also, we controlled participants' observed accuracy for the AI assistant by fixing the model error rate to eliminate the effects of model accuracy on study results; for more discussion of this approach, see (Nourani et al. 2019; Papenmeier, Englebienne, and Seifert 2019). Specifically, the news queue was curated to present one wrong prediction (with equal rate of false positives and false negatives instances) from the AI assistant after every three correct predictions (i.e., true positive and true negative samples), which resulted in a consistent overall 75.00% observed AI accuracy for all participants (participants were not informed or aware of this pattern). Further, as a decision point for trade-offs between clarity and faithfulness of explanations, we performed simple post-processing of model explanations to remove very small features scores in keyword heatmap and normalized articles' attributions scores.

For statistical analysis, inferential tests used one-way independent ANOVAs to compare the conditions for each measure. In the end, we briefly review participants' qualitative feedback to see if they support our quantitative findings.

Human-AI Performance

To answer our first research question, we review and analyze the user performance measure for participants' news reviewing and sharing. We run a between-subject experiment with 40 participants in three primary interface conditions: 1) *Baseline* without any intelligent assistant, 2) Interface with the *AI Assistant*, and 3) Interface with the *XAI-all* Assistant.

Hypothesis 1: Users can share more true news stories with the help of XAI Assistant.

We report the credibility score of participants' shared news as our primary performance measure. Results show the average credibility score is higher than the original news feed (50% credibility) in all three groups that indicates participants' overall ability in news review and their engagement with the task. Participants shared news in *XAI assistant* condition had the highest average of 75.05% (range = [61%, 92%] with SD = 10.06%) credibility and *Baseline* had the least credibility with 68.4% (range = [46%, 88%] with SD = 11.5%) credibility. The data met the assumptions for

parametric testing in all groups with validation checks passing for data normality (Shapiro-Wilk) and homogeneity of variance (Levene's) tests. A significant effect was observed by an ANOVA test with $F(2, 107) = 3.32$ and $p = 0.04$ for the news credibility measure among all three conditions. A post-hoc Tukey test showed borderline significance ($p = 0.050$) between the *XAI assistant* and *Baseline* conditions, with higher news credibility scores for participants in *XAI-all* group compared to *Baseline* group.

We use incredibility score (calculated as *1.0 – credibility score*) of all reported fake news as the secondary performance measures. Similar to credibility scores for shared news, the *XAI* group has the highest average incredibility of reported fake news stories with 73.8% reporting fake news (range = [53%, 100%] with SD = 10.7%). An ANOVA test revealed a significant main effect with $F(2, 107) = 3.78$ and $p = 0.026$. A Tukey post-hoc test showed participants in the *XAI assistant* condition had ($p = 0.019$) reported fake news significantly more than the *Baseline* condition, even though reporting fake news was not the user's primary task during the study.

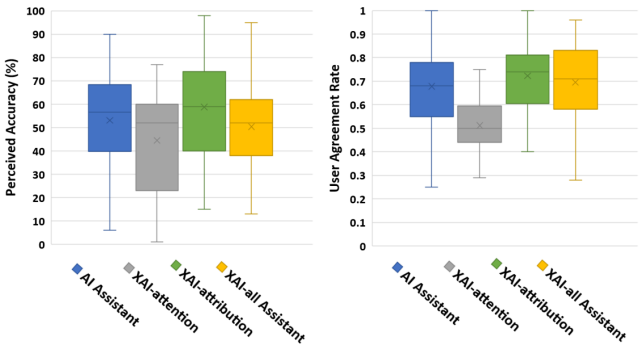
Implications of Results: The study results show that the *XAI assistant* improved user performance compared to the *Baseline* interface without any intelligent agent. However, model explanations did not significantly improve user performance over the *AI assistant* condition. Given the unique design challenges in misinformation detection models, this is a positive indicator that an intelligent agent together with model explanations can potentially improve collaborative human-AI news reviewing.

Mental Model

Our second experiment is designed to answer RQ2 by studying the effects of model explanations on users' mental model. We recruited new participants to ran studies for hypothesis testing through comparison of *AI assistant* condition (as the baseline) with three *XAI assistant* conditions (as treatments) in our interface.

Hypothesis 2: Different types of explanations have different effects on user understanding of intelligent assistants.

Our measure for quantitative evaluation of the mental model is through user prediction task (user guessing of model output). User accuracy in their prediction task was highest ($M = 62.20\%$) in the *XAI-all* group and the worst ($M = 54.65\%$) in the *XAI-attention* group. An ANOVA test detected a significant main effect with $F(3, 149) = 3.16$ and $p = 0.026$ for participants between all four conditions with intelligent assistant. A Tukey post-hoc test yielded a significant difference ($p = 0.017$) between the *XAI-attention* and *XAI-all* groups. However, no significant pairwise difference was detected between the *AI* group and any of *XAI* groups. Note that average user prediction task accuracy in the *XAI-attention* group was lower than the *AI assistance* group, indicating the negative effect of explanations in participants' ability to predict the model output.



(a) User Perceived Accuracy of Intelligent Assistant

(b) User Agreement Rate with Intelligent Assistant

Figure 2: User trust measures for *AI Assistant*, and *XAI Assistants* conditions.

Implications of Results: The results show a significant effect of explanation types on the user mental model based on the user-prediction task measure. However, none of the model explanation conditions improved users' accuracy in prediction. Notably, word-level attention map explanations for news headlines and articles (in the *XAI-attention* condition) had a negative effect on user mental model, potentially due to lower user satisfaction and engagement with the AI assistant. The discrepancy between user prediction task accuracy between the three XAI conditions indicates that not all explanations are informative or meaningful for end-users to be able to predict model behavior.

Trust and Reliance

To address RQ3, we review and analyze user trust and reliance measures in our experiments.

Hypothesis 3: Users have higher perceived accuracy in XAI assistant compared to AI Assistant.

Our primary measures for user trust in the AI and XAI assistant is the participants' perceived accuracy of the intelligent assistant. Figure 2a shows a box-plot of participants' perceived accuracy of the AI and three XAI assistants conditions. The results show participants had the highest rate of perceived accuracy in the *XAI-attribution* group (with the visualization of news feature attribution) and lowest in the *XAI-attention* group (with the heatmap of word feature attribution) on average. Using an ANOVA test, we found a significant difference ($F(3, 155) = 2.86$ and $p = 0.039$) between perceived accuracy in the four groups. For pair analysis, a post-hoc Tukey test revealed participants' perceived model accuracy of the *XAI-attribution* condition ($M = 58.70\%$) was significantly ($p = 0.024$) higher than *XAI-attention* ($M = 45.55\%$). Interestingly, participants in the *XAI-all* group responded with lower perceived accuracy ($M = 50.38\%$) compared to AI Assistant ($M = 53.05\%$) with no explanation.

Hypothesis 4: Users will agree more with XAI assistant predictions compared to AI assistant.

We measure user reliance on algorithms via the user agreement rate with AI and XAI assistants predictions. Figure 2b presents results for participants agreement rate with the *AI assistant* and three *XAI assistant* groups. Overall, participants had near 0.70 agreement rate with model prediction in all groups except for the *XAI-attention* group with 0.51 agreement rate. We observed a significant main effect using an ANOVA test with $F(3, 149) = 16.44$ and $p < 0.001$ for participants agreement rate with intelligent assistants prediction. From the pairwise Tukey post-hoc analysis, participants had a significantly lower agreement rate in the *XAI-attention* group compared to all three other groups ($p < 0.001$ for all pairwise comparisons). Similar to participants' perceived accuracy, tests did not detect a significant increase in user agreement from model explanations.

Implications of Results: The study results indicate that model explanations helped users to adjust their trust and reliance on the intelligent assistant. We did not observe improvements in user trust or reliance for the XAI assistants over the AI assistant. In fact, participants actually lost trust in the *XAI-attention* assistant when—despite their initial expectations—they found the system was detecting fake news only based on news keywords. This could be considered an appropriate result given the limitations of the model logic. The lower user trust in the *XAI-attention* condition coincides with participants' mental model results and might suggest the effectiveness of explanations in helping users avoid overtrusting the intelligent assistant in cases when model logic may not be optimal or meaningful based on human logic.

Qualitative Feedback

Reviewing participants' written feedback in the post-study survey reflects their reasoning about AI assistant that provides further insight into participants' mental models of the AI/XAI assistants. Participants answered two descriptive questions regarding their mental model of the AI assistant's reasoning ("*How do you describe this AI's reasoning to find fake news?*") and AI assistant's limitations ("*In your opinion, what are the biggest limitations of this AI fake news detector?*"). Two authors separately reviewed participants' qualitative feedback and performed open coding to extract themes in participants' notes and comments. Two authors coded participants' free response questions to identify salient themes. Over three sessions of coding and discussion, they identified 19 codes with an inter-rater reliability of 0.82. We then used these codes to create three main categories of responses: AI reasoning, AI limitations, and participant-strategy.

Regarding participant mental models of AI assistants, we observed that explanations clearly improved their understanding of AI reasoning. On average, 63.5% of participants in the *XAI-attention* and 52.8% of in the *XAI-all* group pointed out the importance of keywords in the news; example comments include "*I think it looked for certain key words*" and "*The AI compares relevant phrases in the headline to relevant keywords in the supporting stories.*" In contrast, only 17.9% of participants in the *AI assistant* condition had expressed such understanding. We also found 62.0% participants in the *XAI-attribution* group mentioned related articles and their sources as key features for AI reasoning compared to 31.7% in the *XAI-attention* group. For example, one participant in this group commented "*It tries to pull related articles from the web to prove or disprove the headline*", and another participant said "*I think it went by how many articles below matched the news.*

We also found interesting feedback on participants' subjective opinions on the limitations of the AI assistant. We saw a clear theme in responses to the need for common sense to distinguish fake and true news. On average, from 20% of participants in all conditions (except *XAI-all* with 11.1%), we received comments such as "*it doesn't have human judgment*", "*I guess they will not see common sense*", and "*The AI doesn't have the experience that a real person has in dealing with the fake news out there.*". Also, participants in *XAI-attention* group paid more attention to the quality and combination of articles in each news story with 43.1% of them expressing comments like "*AI doesn't have enough information*" and "*It doesn't see multiple sides of the story*" compared to other conditions with the average of 19.3%. Additionally, 27.2% of all participants expressed concern about AI ability in understanding the context of the news or recognizing sarcasm. For instance, one said "*I think it can't detect sarcasm satire or parody so it has some limitations*" and another mentioned that "*The AI isn't able to understand the context of the text. It's not able to actually understand the story or [its] plausibility.*".

Challenges participants encountered in learning the model behavior was also reflected in 13.5% of participants' comments in all groups, for example one said:

> *I said [to myself] twice that I thought I understood how it worked but when asked to predict the AI's inference about a given headline in the last portion of the study I believe I only matched one out of four so maybe I didn't understand anything that well.*

Overall, the qualitative user feedback complement the quantitative findings in showing how explanations helped participants to observe model limitations and adjust their trust and reliance accordingly.

Interactions Between XAI Measures

In this section, we summarize different implications of our study results for machine learning explanations and fake news detection. Following Hoffman et al.'s (2018) conceptual model (Figure 3), we look for correlations between our measurements of user engagement, mental model, performance, and trust to investigate the interplay between these factors.

User Expectations of AI Assistant

We first analyze the relation between user expectations of AI before the study and their perceived algorithm accuracy after the study. Research shows that various external and internal factors can interact with user trust, with examples including

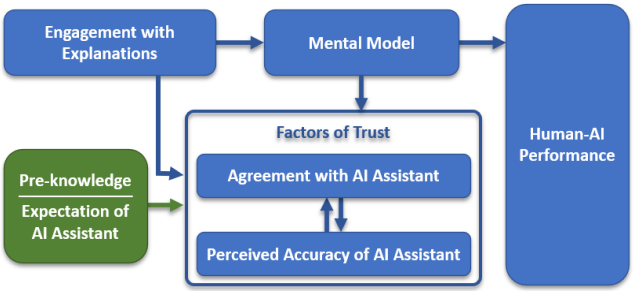


Figure 3: Conceptual model of relationships among user engagement, mental model, trust, and human-AI performance in XAI systems. Figure created based on a model of the "process of explaining" in XAI context from (Hoffman et al. 2018).

user pre-knowledge (Horne et al. 2019), model stated performance (Yin, Wortman Vaughan, and Wallach 2019), and model observed performance (Nourani et al. 2019). In the pre-study questionnaire, we measured 1) participant expectation of AI assistant accuracy and (with "If you had an Artificial Intelligence (AI) algorithm to review your daily news for fake news detection, what would be your expectation of AI accuracy to do a good job?" question) and 2) participant estimation of fake news rate in media (with "In your experience, what percentage of news that you read daily is false news? e.g., fake news, hoax, rumors, made up stories, misinformation" question).

As expected, a Pearson test shows a positive correlation ($r = 0.223$, $p = 0.005$) between participants' perceived accuracy at the end of the study and their initial expectation of AI accuracy. Regarding participants' expectations of fake news occurrence in daily news, we expected to see more user engagement for participants with higher anticipation of fake news. However, we surprisingly found that participants expectation of fake news occurrence has a negative correlation with their engagement with AI assistant ($r = -0.189$, $p = 0.018$). This could be due to participants underestimating the AI assistant or choosing their intuition rather than model suggestions.

Engagement with Intelligent Assistants

As an objective measure of user engagement with intelligent assistants, we consider total continued usage based on the frequency of user interactions (clicks count) with the AI and XAI assistant predictions. Overall average results show that participants in the *XAI-all* group had the highest engagement rate with the XAI assistant (0.95 prediction check rate) for shared or reported news stories. An ANOVA test of user engagement with the AI assistant found a significant difference with $F(3, 156) = 2.773$ and $p = 0.046$ between conditions, and a Tukey post-hoc test shows participants were significantly ($p = 0.034$) more engaged in the *XAI-all* condition compared to *XAI-attention* condition.

The conceptual model of the process of explaining (Hoffman et al. 2018) suggests that explanations in XAI system revise mental model and can engender appropriate trust, see Figure 3. To test the interplay between user engagement and their mental model of XAI assistants, we performed a bivariate Pearson correlation test between user engagement rate and prediction task accuracy as the mental model measure. Despite the initial hypotheses, a Pearson correlation did not show a positive relation between engagement and mental model ($r = 0.099, p = 0.215$). This could be due to the narrow scope of mental-model measurement in our study being limited to the user prediction task (model predictability for users). However, user engagement had a significant positive correlation ($r = 0.228$, $p < 0.001$) with user agreement with the intelligent assistant. This shows as more participants got involved with the AI or XAI predictions, the more they agreed with its predictions.

Mental Model Affecting Performance and Trust

Next, we analyze how users' mental model interacts with trust and human-AI performance. A Pearson test between users' prediction task accuracy (mental model measure) and perceived accuracy of AI assistant (our first user trust measure) showed a positive significant correlation ($r = 0.212$, $p = 0.008$). A correlation test between user prediction accuracy and user agreement with the AI assistant (our second user-trust measure) also showed a significant positive correlation ($r = 0.280$, $p < 0.001$) between participants' mental model and trust. Positive correlations of the mental model measure with both trust measures demonstrate the relation between the intelligent agent's predictability and trust.

As hypothesized, user prediction task accuracy was positively correlated with credibility of shared news ($r = 0.305$, $p < 0.001$) as well as incredibility of reported fake news ($r = 0.283$, $p < 0.001$). This finding suggests users with a more accurate mental model could better guess model failure cases, and by avoiding those cases, they could improve their performance.

Interactions Between Trust Measures

Another interesting finding from our study is that we observed interactions between multiple measures of user trust. Previous research studies have utilized various independent trust measures such as perceived algorithm performance (Nourani et al. 2019), perception of control over the system (Kocielnik, Amershi, and Bennett 2019), and the rate of users' agreement with an algorithm's recommendations (Yin, Wortman Vaughan, and Wallach 2019). In our studies, we measured two different trust factors to examine how they may interact. A Pearson correlation test between the two trust measures shows a significant positive correlation between the perceived accuracy and user agreement rate ($r = 0.482$, $p < 0.001$). This positive correlation suggests that as users feel more confident about AI competence, they tend to agree more with its predictions.

Conclusion and Future Work

In our research, we evaluated model explanations from multiple models as part of an ensemble approach for fake news detection. This approach allowed us to study how different types of explanations affect users in fake news detection.

In conclusion, our research revealed multiple challenges in designing effective XAI systems in the fake news detection domain. In particular, we observe challenges rising from the inherent difference between models' feature learning (word-level features in our case) and human understanding of news and information. Overall, users' interaction with the AI and XAI assistants affected their performance, mental model, and trust. However, model explanations in our studies did not improve task performance or increase user trust and mental model. Instead, the quantitative results and qualitative feedback indicate that explanations helped users' to build an appropriate mental model of intelligent assistants and adjust their trust accordingly given their perception of limitations of the models. For example, participants in the *XAI-attention* group that was significantly less successful in guessing model outputs also showed significantly lower trust (in both trust measures) compared to the *XAI-all* condition. Likewise, reviewing user engagement results showed that *XAI-attention* explanations were not appreciated by the users. Therefore, we conclude that improving transparency of the model helped users to appropriately avoid overtrusting the fake news detector when they found the AI reasoning was not trustworthy or simply explanations were nonsensical.

We also recognize a few limitations in our experiments that could become more clear in future work. First, fake news detection is a challenging task for machine learning models, and the low model performance could have affected participants' perception of AI assistant competence and therefore study results. Similarly, possible biases in news labels from training set could have affected models and consequently participants' perception of model limitations.

Additionally, given the complexity of the user task in our study design and possible participants' biases toward news samples, our study's sample size in each condition may be a limitation in our experiments. Our experiments studied multiple types of NLP explanations, however, future research is needed to assess the effectiveness of other types of explanations, such as knowledge graphs and multi-modal evidence retrieval for users in fake news detection task.

Acknowledgments

This work is in part supported by DARPA grant N66001-17-2-4031, NSF award #1900767, and NSF award #1900990. The views and conclusions in this paper are those of the authors and should not be interpreted as representing any funding agencies. We would also like to thank our Mechanical Turk participants.

References

Chen, T., and Guestrin, C. 2016. Xgboost: A scalable tree boosting system. In *KDD*. ACM.

Du, M.; Liu, N.; and Hu, X. 2019. Techniques for interpretable machine learning. *Communications of the ACM* 63(1):68–77.

Eslami, M.; Rickman, A.; Vaccaro, K.; Aleyasen, A.; Vuong, A.; Karahalios, K.; Hamilton, K.; and Sandvig, C. 2015. I always assumed that i wasn't really that close to [her]: Reasoning about invisible algorithms in news feeds. In *CHI*. ACM.

Eslami, M.; Krishna Kumaran, S. R.; Sandvig, C.; and Karahalios, K. 2018. Communicating algorithmic process in online behavioral advertising. In *CHI*. ACM.

Gunning, D. 2017. Explainable artificial intelligence (XAI). *Defense Advanced Research Projects Agency (DARPA)*.

Hechtlinger, Y. 2016. Interpretation of prediction models using the input gradient. *NIPS*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2018. Metrics for explainable ai: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Horne, B. D.; Nevo, D.; O'Donovan, J.; Cho, J.-H.; and Adalı, S. 2019. Rating reliability and bias in news articles: Does ai assistance help everyone? In *ICWSM*.

Huang, Z.; Xu, W.; and Yu, K. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Kocielnik, R.; Amershi, S.; and Bennett, P. N. 2019. Will you accept an imperfect AI?: Exploring designs for adjusting end-user expectations of AI systems. In *CHI*. ACM.

Kumar, A., and Rastogi, R. 2019. Attentional recurrent neural networks for sentence classification. In *Innovations in Infrastructure*. Springer.

Li, J.; Monroe, W.; and Jurafsky, D. 2016. Understanding neural networks through representation erasure. *arXiv preprint arXiv:1612.08220*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *NIPS*.

Mohseni, S.; Ragan, E.; and Hu, X. 2019. Open issues in combating fake news: Interpretability as an opportunity. *arXiv preprint arXiv:1904.03016*.

Mohseni, S.; Zarei, N.; and Ragan, E. D. 2019. A multidisciplinary survey and framework for design and evaluation of explainable ai systems. *arXiv preprint arXiv:1811.11839*.

Murdoch, W. J.; Liu, P. J.; and Yu, B. 2018. Beyond word importance: Contextual decomposition to extract interactions from lstms. *ICLR*.

Nguyen, A. T.; Kharosekar, A.; Krishnan, S.; Krishnan, S.; Tate, E.; Wallace, B. C.; and Lease, M. 2018. Believe it or not: Designing a human-ai partnership for mixed-initiative fact-checking. In *UIST*. ACM.

Nourani, M.; Kabir, S.; Mohseni, S.; and Ragan, E. D. 2019. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *HCOMP*.

Papenmeier, A.; Englebienne, G.; and Seifert, C. 2019. How model accuracy and explanation fidelity influence user trust. *arXiv preprint arXiv:1907.12652*.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global vectors for word representation. In *EMNLP*. ACL.

Popat, K.; Mukherjee, S.; Yates, A.; and Weikum, G. 2018. DeClarE: Debunking fake news and false claims using evidence-aware deep learning. In *EMNLP*. ACL.

Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; and Stein, B. 2017. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*.

Poursabzi-Sangdeh, F.; Goldstein, D. G.; Hofman, J. M.; Vaughan, J. W.; and Wallach, H. 2018. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*.

Rader, E.; Cotter, K.; and Cho, J. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *CHI*. ACM.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *KDD*. ACM.

Rubin, V. L., and Lukoianova, T. 2015. Truth and deception at the rhetorical structure level. *JASIST*.

Shu, K.; Cui, L.; Wang, S.; Lee, D.; and Liu, H. 2019. defend: Explainable fake news detection. In *KDD*. ACM.

Shu, K.; Wang, S.; and Liu, H. 2017. Exploiting tri-relationship for fake news detection. *arXiv preprint arXiv:1712.07709*.

ter Hoeve, M.; Heruer, M.; Odijk, D.; Schuth, A.; and de Rijke, M. 2017. Do news consumers want explanations for personalized news rankings. In *FATREC Workshop on Responsible Recommendation Proceedings*.

Trielli, D., and Diakopoulos, N. 2019. Search as news curator: The role of google in shaping attention to news information. In *CHI*. ACM.

Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; and Hovy, E. 2016. Hierarchical attention networks for document classification. In *NAACL-HLT*, 1480–1489.

Yang, F.; Pentyala, S. K.; Mohseni, S.; Du, M.; Yuan, H.; Linder, R.; Ragan, E. D.; Ji, S.; and Hu, X. B. 2019. Xfake: Explainable fake news detector with visualizations. In *WWW*. ACM.

Yin, M.; Wortman Vaughan, J.; and Wallach, H. 2019. Understanding the effect of accuracy on trust in machine learning models. In *CHI*. ACM.