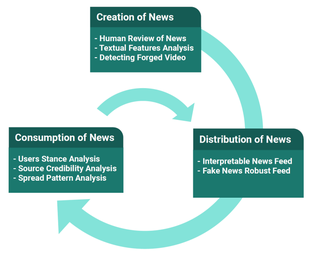
# Machine Learning Explanations to Prevent Overtrust in Fake News Detection

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]Department of Computer Science and Engineering, Texas A&M University, USA
[2]Department of Computer & Information Science & Engineering, University of Florida, USA

## INTRODUCTION AND MOTIVATION




⇒ **News feed algorithms function similar to decision-making algorithms, as users are exposed to algorithmically selected content.**

⇒ **Evidence show news feed could potentially lead to unintentional large-scale propagation of false and fabricated information.**

## ALGORITHMIC TRANSPARENCY IN SOCIAL MEDIA

**We posit Explainable AI (XAI) systems can be a solution to fair automated news review by self-explaining the reasoning behind model predictions.**

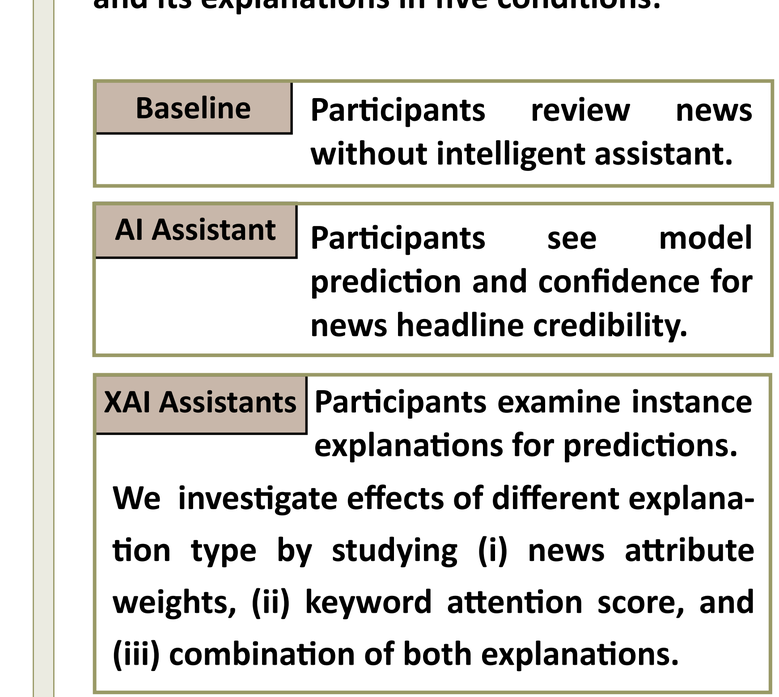
**We study the effects of algorithmic transparency in fake news detection, specifically we are interested to learn:**

⇒ **Do model explanations help end-users to share more credible news?**

⇒ **Do model explanations improve user mental model?**

⇒ **How do different factors of user trust get affected by AI explanations?**

**For a deeper understanding of Explainable AI systems, we study relationships among multiple external (green) and internal (blue) human factors in Human-AI interactions based on Hoffman et al's conceptual model of the "process of explaining" in XAI context.**




## INTERPRETABLE FAKE NEWS DETECTION SYSTEM

**We built a news review system with interpretable fake new detection to serve our research goals and human-subject studies:**

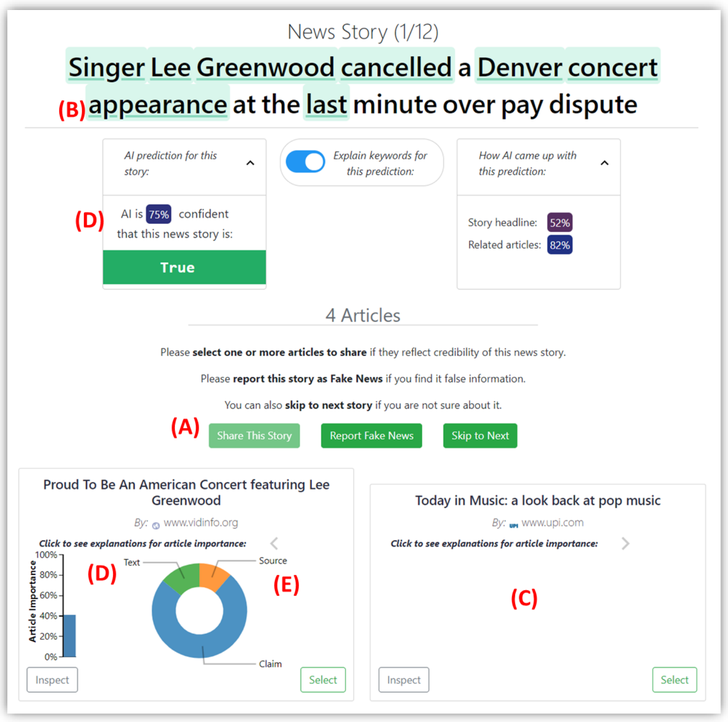
✓ **We created a dataset of news stories from two sources: a) 4638 news claims and labels from Snopes fact-checking website each with b) 16 related news articles crawled from web using top Google search results.**

✓ **An ensemble of four classifiers (D) are trained on our training set for fake news detection:**

⇒ **Two models take news claim as input. We used a Bi-LSTM network (B) and a hierarchical attention network (D) that learns relation between articles and claim.**

⇒ **Two models are trained on supporting articles. We used a mimic learning framework to train XGBoost (E) student model and a Bi-LSTM network trained on news articles.**

✓ **We designed an online interface for users to review a queue of news stories, share true news (A) for other users, and report fake news. The interface, with build-in XAI Assistant (D), shows the news headline on the top (B) followed by a list of related articles (C) underneath.**




## STUDY DESIGN

**We conducted a human-subjects study with 220 AMT participants for controlled comparison of elements of the fake news detector and its explanations in five conditions:**

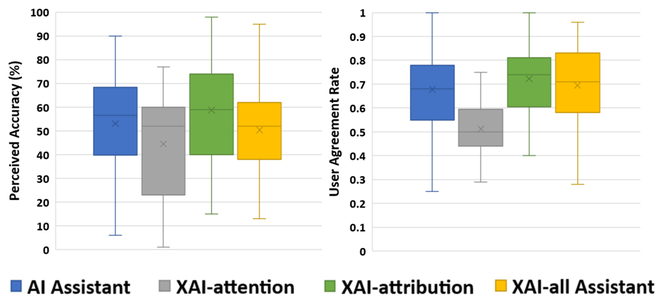
| | |
|---|---|
| **Baseline** | **Participants review news without intelligent assistant.** |
| **AI Assistant** | **Participants see model prediction and confidence for news headline credibility.** |
| **XAI Assistants** | **Participants examine instance explanations for predictions.** |

**We investigate effects of different explanation type by studying (i) news attribute weights, (ii) keyword attention score, and (iii) combination of both explanations.**

## USER TRUST AND RELIANCE

**We measured user trust with *subjective user perceived accuracy* and *user agreement rate* with the intelligent assistant.**

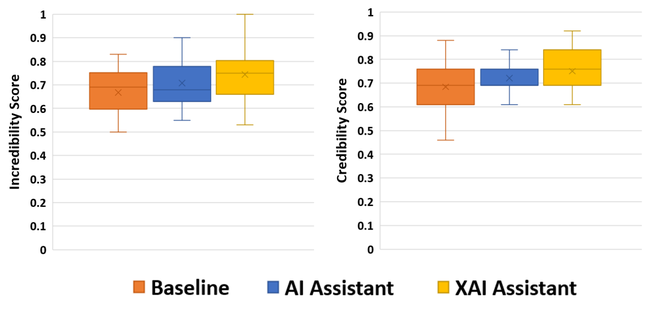
**For both trust measures, the XAI Assistant with news attribute explanations was significantly more trusted compared to keyword attention explanations.**




## NEWS CREDIBILITY

**We measured *credibility of shared news* and *incredibility of reported news* as our performance metrics.**

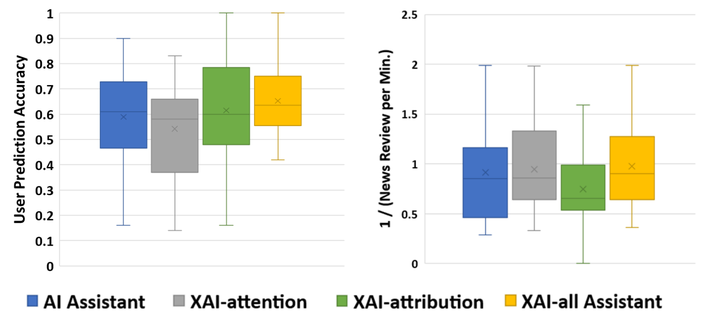
**We observed statistically significant difference between Baseline and XAI Assistant conditions in which participants from XAI group outperformed the other two groups in both performance measures.**




## USER MENTAL MODEL AND COGNITIVE LOAD

**Our measure for evaluating users mental model was *user accuracy in guessing* the model output during final four news reviews.**

**Statistical tests on participants data revealed that XAI group had significantly better mental model compared to keyword attention alone, however, the news review duration was significantly longer for this group.**




## ACKNOWLEDGMENT

**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.**