Trust Evolution Over Time in Explainable AI for Fake News Detection

Sina Mohseni
Fan Yang
Shiva Pentyala
Mengnan Du
Yi Liu
Nic Lupfer
Xia Hu
Shuiwang Ji
Texas A&M University
College Station, TX
sina.mohseni@tamu.edu
nacoyang@tamu.edu
pk123@tamu.edu
dumengnan@tamu.edu
yiliu@tamu.edu
nlupfer@tamu.edu
hu@cse.tamu.edu
sji@tamu.edu

Eric D. Ragan
University of Florida
Gainesville, FL
eragan@ufl.edu

Author Keywords

Explainable AI; User Trust; Perceived Accuracy; Expectation of Accuracy; Fake News Detection.

CHI'20,, April 25–30, 2020, Honolulu, HI, USA
ACM 978-1-4503-6819-3/20/04.
https://doi.org/10.1145/3334480.XXXXXXX

Abstract

The need for interpretable and accountable intelligent systems is strong as artificial intelligence (AI) becomes more prevalent in human life. We study the effects of interpretability on user's trust in an AI assistant tool designed for fake news detection. In our study, we expose participants to different types of AI and Explainable AI (XAI) assistants, measure their perceived accuracy of algorithm, and cluster user trust changes over time into five types of trust evolution. We present quantitative results and analysis from human-subject studies and discuss our findings regarding how model explanations affect on user trust evolution over time.

CCS Concepts

•**Human-centered computing** → **Human computer interaction (HCI);** User studies;

Introduction and Background

There has been a surge in AI-based products and services to process large data, enable autonomy, and enhance end-users experience in recent years. However, lack of transparency and specifiability in advanced AI algorithms can potentially result in unfair and unsafe decision-making. In this regard, growing attention on the transparency of algorithmic decision-making systems is looking for solutions to eliminate possible biases and errors in AI-based systems.

Explainable AI (XAI) systems are proposed as an opportunity for intelligible and predictable AI via providing explanations such as “why” the prediction is made and “how” the AI works for end-users.

Studying user trust in intelligent systems is an important topic and is drawing more attention with the increasing end-user interactions with AI-infused systems. In the context of XAI, research investigates how different types and amount of explanations affect user trust and reliance in the intelligent systems [4]. Related to this, various methods and measures have been proposed to evaluate different factors of trust, including perception of algorithm performance [5], perception of control over system [3], perception of algorithm transparency [1], and user agreement rate with the algorithm recommendation. For example, Yin et al. [6] used both user agreement rate with the algorithm and decision switch rate after seeing algorithm recommendation as to their trust measures.

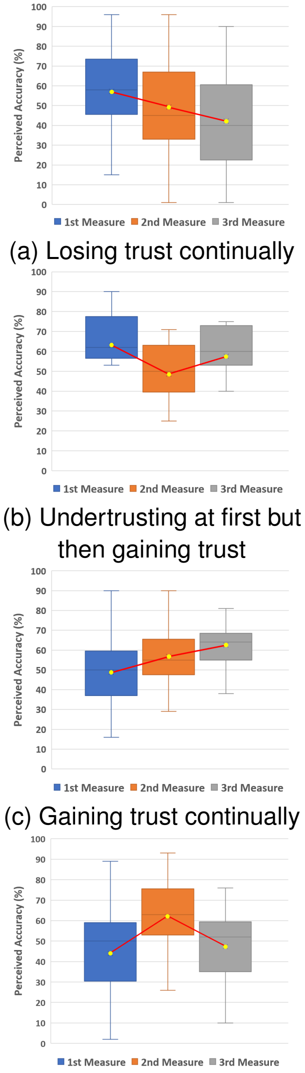
We contribute to research on user trust in intelligent systems with a series of human-subject experiments to study *user trust evolution over time* in a case study with an explainable fake news detector. To this end, this paper presents a crowdsourced study with an interactive system designed for reviewing and sharing news stories and articles. Our system provides a built-in interpretable fake news detector as an AI assistant for non-expert end-users. We present and discuss our study results and analysis to answer: *How does user trust evolve over time at presence of explanations?*

System Design

To answer our research questions, we first build a news review interface with a build-in AI assistant and then run human-subject experiments for hypothesis testing. We crawled our news dataset from news headlines in Snopes (with ground truth) and supporting articles from the top 16 Google search results.

Explainable Interface: We design an interface in which users can review news headlines and choose news stories to share with other users. Figure 1 shows our news review interface that enables the main user task and interactions with the AI assistant. Users review news stories one-by-one and decide if the 1) story is true to be shared with other users, or 2) story is fake news to be reported, or 3) they want to skip to the next story for any reason. The interface shows a list of related articles for each news story that provides context about the news headline. A fake news detection assistant is designed in our system which provides predictions about the news veracity. Different types of instant explanations including keyword importance score, articles score, and source importance score are embedded in the interface as instance explanation.

Interpretable Models: We implement an ensemble of four classifiers for fake news detection. Our first model is a LSTM network with a self-attention layer trained on news headlines that provides explanations in form of salient words for news headlines. The second model performs fake news detection based on article set representation constructed using hierarchical attention at sentence level and article-level producing attribution score explanations. Our third model uses a knowledge distillation approach to approximate a deep architecture with a random forest to generate attribute importance (i.e., news claims, articles, and news sources) for each news classification. Our last model uses crawled related articles captured by a bidirectional LSTM module and explains its prediction with salient words.



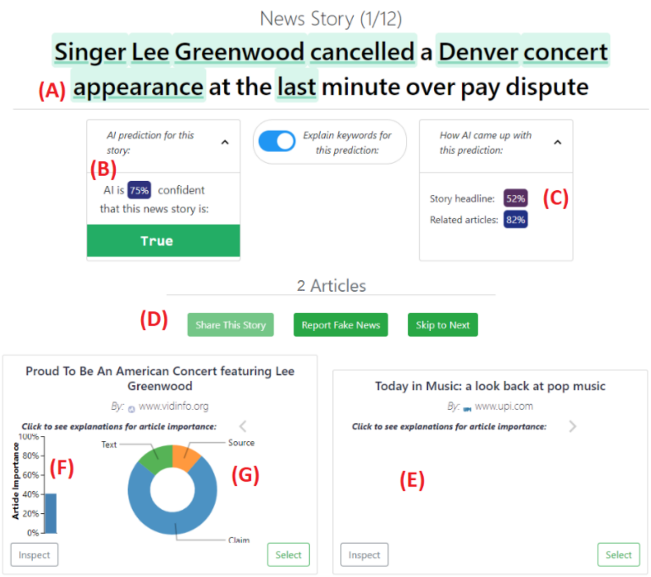
(a) Losing trust continually

(b) Undertrusting at first but then gaining trust

(c) Gaining trust continually

(d) Overtrusting at first but correcting over time

Figure 2: Four profiles of user trust evolution in time.


News Story (1/12)

Singer Lee Greenwood cancelled a Denver concert appearance at the last minute over pay dispute

AI prediction for this story:

AI is 75% confident that this news story is:

True

Explain keywords for this prediction:

How AI came up with this prediction:

Story headline: 52%

Related articles: 82%

2 Articles

Proud To Be An American Concert featuring Lee Greenwood

By: www.vidinfo.org

Click to see explanations for article importance:

Today in Music: a look back at pop music

By: www.upi.com

Click to see explanations for article importance:

Figure 1: Our explainable user interface. A) News headline with a heatmap explanation. B) Model prediction and confidence. C) Prediction confidence for different inputs. D) User choices to share, report, or skip the news story. E) Supporting news articles for the headline. F) Bar chart visualization for article importance score. G) Donut chart visualization for article attribution scores.

Experiment and Results

We designed and ran a between-subject studies with 160 participants from Amazon Mechanical Turk. Participants were recruited only from *master* workers with above 90% acceptance rate. In all experiments, news stories were sorted in the news queue in order to show one fake news for every true news. Also, the accuracy of the AI assistant was controlled such that it made a wrong detection (with equal rate of false positives and false negatives instances) after every three correct predictions (i.e., true positive and true negative samples), resulting in an overall 66.6% observed accuracy for participants.

Study Conditions: For our experiments, we treat groups of participants with four different interface conditions. Specifically, we create an AI condition (Baseline without explanations) and two XAI conditions (with explanations) from our ensemble of fake news detectors as follows: 1) AI condition only shows model predictions (Figure 1-(B and C)), 2) XAI-attention condition presents model predictions explained by a heatmap of word importance for news headline and articles (Figure 1-(B, C, and A)), 3) XAI-attribute condition shows model predictions explained by a donut chart of supporting articles attribution score (Figure 1-(B, C, F, and G).

Trust Measure: For our trust measures, we measure participants' perception of AI accuracy in mid-task and post questionnaires. We measure user perceived accuracy twice during the study (at 1/3 and 2/3 study progress) and once at the post-study questionnaire by asking user "What was the accuracy of the AI fake news detection?". For all measures, we collect participants' feedback in the range of 0 to 100% using a slider bar with the step size of 1.

Trust Evolution Over Time

We analyze our repeated trust measurements during the study to investigate how user perception of algorithmic performance evolved over time. We measured participants' subjective accuracy of AI and XAI assistants three times (in intervals of four news reviews) during the studies, aiming to record potential trends of user trust evolution in intelligent assistant. For the analysis, we first hypothesized and clustered (rule-based clustering) trends of user trust changing over time based on the following types of changes: **Type 1)** the user is losing trust continually, **Type 2)** user trust undershoots, **Type 3)** user gains trust continually, **Type 4)** user trust overshoots, **Type 5)** user trust level remains constant.

Study Procedure: Study starts with participants accepting the information sheet (IRB approved) and filling the per-study questionnaire. Participants are prompt to share 12 true news (regardless of number of reporting fake news or skip news) divided into 3-sets of news review each followed by a mini mid-study questionnaire. Study ends with a post-study survey of Likert-scale questions and open ended descriptive questions.

Sidebar 1: Study Procedure

Acknowledgements

This work is in part supported by DARPA grant N66001-17-2-4031. 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.

Figure 2 shows our clustering results for overtime trust measurements into four trust evolution types. Overall most 36.3% of participants over-trusted the AI assistant (Type 4) initially, 23.6% gained trust continually (Type 3), 21.0% lost trust continually, 10.8% did not change their subjective trust feedback (Type 5), and 8.3% had trust undershoot during the task (Type 2). A Pearson Chi-square test shows a significant correlation (p = 0.033) between explanation conditions (the *AI* and two *XAI* conditions) and user trust types. The majority of participants from the *XAI-attention* condition had overshoot in their second perceived accuracy measurement. In comparison, more participants from the *AI* and *XAI-attribute* conditions were continuously gaining trust in the system.

Conclusion and Future Work

In our experiments, we studied how model explanations affect users perceived accuracy of algorithms over time. Analysis of over time trust measures revealed valuable insights on user behavior when interacting with the intelligent systems. Unlike Holliday et al. [1], we focused on quantitative analysis of user feedback on their perceived accuracy for insights. We clustered user trust evolution over time into five types (see Figure 2), and a Chi-square test unveiled significant effect of machine learning explanations on user trust changes. Following related research on user trust in intelligent systems (e.g., [6, 2]), we conclude that AI transparency and machine learning explanations do not necessarily improve user trust; instead, transparency empowers the user to build appropriate trust in the system. Further, explanations type also affect on how user trust would evolve to its stable state.

In future work, we plan to analyze the correlation between multiple trust factors (e.g., user agreement with AI and user perceived accuracy) and their relation to task performance. We expect our studies to lead to modeling user behavior based on interactions and trust factors to calibrate user trust in the AI system for improving Human-AI collaboration performance. For example, adjusting AI/XAI recommendations and explanation types based on user trust level during user task could benefit overall Human-AI collaboration performance. We also plan to conduct qualitative analysis for more insights on “why” explanations types resulted in different user trust journeys.

REFERENCES

[1] Daniel Holliday, Stephanie Wilson, and Simone Stumpf. 2016. User trust in intelligent systems: A journey over time. In *CHI*. ACM.

[2] René F Kizilcec. 2016. How much information?: Effects of transparency on trust in an algorithmic interface. In *CHI*. ACM.

[3] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. In *CHI*. ACM, 411.

[4] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2019. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *arXiv preprint arXiv:1811.11839* (2019).

[5] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D Ragan. 2019. The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems. In *HCOMP*.

[6] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *CHI*. ACM.