Received: 21 June 2021 | Revised: 24 September 2021 | Accepted: 25 October 2021

DOI: 10.1002/ail2.49

LETTER

WILEY

How level of explanation detail affects human performance in interpretable intelligent systems: A study on explainable fact checking

Rhema Linder[1] | **Sina Mohseni**[2] | **Fan Yang**[2] | **Shiva K. Pentyala**[2] | **Eric D. Ragan**[3] | **Xia Ben Hu**[4]

[1]Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, Tennessee, USA

[2]Department of Computer Science, Texas A&M University, College Station, Texas, USA

[3]Department of Computer, Rice University, Houston, Texas, USA

[4]Department of Computer and Information Science and Engineering, University of Florida, Gainesville, Florida, USA

Correspondence
Rhema Linder, Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN.
Email: rlinder@utk.edu

Funding information
DARPA, Grant/Award Number: N66001-17-2-4031

Abstract

Explainable artificial intelligence (XAI) systems aim to provide users with information to help them better understand computational models and reason about why outputs were generated. However, there are many different ways an XAI interface might present explanations, which makes designing an appropriate and effective interface an important and challenging task. Our work investigates how different types and amounts of explanatory information affect user ability to utilize explanations to understand system behavior and improve task performance. The presented research employs a system for detecting the truthfulness of news statements. In a controlled experiment, participants were tasked with using the system to assess news statements as well as to learn to predict the output of the AI. Our experiment compares various levels of explanatory information to contribute empirical data about how explanation detail can influence utility. The results show that more explanation information improves participant understanding of AI models, but the benefits come at the cost of time and attention needed to make sense of the explanation.

KEYWORDS

explainable artificial intelligence, human-computer interaction, machine learning, transparency

INTRODUCTION

Artificial intelligence (AI) and machine learning (ML) provide powerful tools for assisting with data analysis and human decision-making across a variety of application areas. Unfortunately, computational models can be complicated for nonexperts to understand, and many end users may be reluctant to trust and rely on intelligent systems they do not understand. It is not surprising that experts have long argued for transparency in computing,[1,2] and recent advances in AI and deep learning methods have been accompanied by growing concerns about black-box computing.

To address such issues, researchers have expressed interest in *explainable artificial intelligence* (XAI)[3] and *interpretable machine learning* solutions.[4] XAI systems aim to provide users with information to help better understand how computational models work and why their outputs are generated. Ideally, a better understanding of machine

intelligence could allow users to identify and correct errors in the models, or to make informed decisions about when to trust a system's output and when to be more skeptical. Such improved understanding has the potential to facilitate better decision-making for real-world tasks.

The challenge, however, is in how to design effective and human-understandable explanations. Designers have near limitless options for presenting meaningful and helpful explanations to users. Explanations can come in different formats. They may be textual,[5,6] visual,[7,8] or numerical.[9,10] Explanations may vary greatly in the level of detail they present about their underlying model and rationale about specific outputs.

Our research focuses on studying trade-offs in the levels of explanation details for XAI systems for human judgment tasks. Intuitively, explanations need to provide enough information to enable understanding and support human decision-making. On the other hand, too much information may be confusing, distracting, or overwhelming to users. We present a controlled experiment to study how different amounts of explanation affect human decision-making and human understanding of a classifier. Our experiment is grounded in the context of an explainable classification system for fact-checking news statements. Determining the validity of news statements is a difficult task of interest to the general public, journalists, and politicians alike.[11-13] For our research, we designed an explanation interface that utilizes past news statements for automatic classification with integrated XAI interface components. Our study puts participants in the role of assessing the accuracy of news statements with the assistance of our classifier. The experiment compares participant performance for different levels and types of explanation provided to participants while reviewing and assessing news statements.

In this paper, we describe the experiment, the XAI system, and the study results based on quantitative differences in time and accuracy across conditions. We also demonstrate how participants viewed their experience based on their qualitative responses. Our findings show the XAI system improved their performance in some cases, but that presenting too much information can negatively impact human judgment. We discuss these results and make recommendations for future research in XAI.

RELATED WORK

We review related research to the explainable machine learning systems. Also, because our study and XAI system relies on a fact-check detection task, we discuss relevant background on news-checking methods and services.

Overview of explainable machine learning

Explainable machine learning is a class of machine learning algorithms that generate interpretable explanations of behavior and predictions. Machine-generated explanations are complementary information designed to help users understand various complex machine learning decision-making processes. Research show machine learning interpretability could result in improving user experience, trust, and performance in different applications.[14,15] Doshi-Velez and Kim consider the need for explanations to be derived from an incompleteness in the problem formalization.[16] They discuss how domain experts would better trust the system with explanations to verify scientific findings, confirm economic predictions, and justify fair and ethical decision-making. Machine-learning designers also benefit from explanations to adjust and optimize models for the right objectives.[17,18]

A recent review by Mohseni et al[19] explains how different research communities take different approaches and adopt different priorities in improving explainability. Common goals of machine-learning explanation include describing the entire machine-learning model or explaining individual instances.[20] Model explanations, sometimes called global explanations, aim to directly represent or summarize the entire model using interpretable explanations. However, in many cases, machine-learning product end users do not have access to the entire machine-learning model, leading researchers to present relatively simple visual[21-23] and verbal[5,24] representations of model reasoning. Visual analytics tools also help machine-learning engineers to visualize model parameters[25,26] and review the training process[27] for better design and evaluation of systems. In contrast to model explanations, instance explanations, sometimes called local explanations, describe prediction details for individual input instances. In practice, many implementations of instance explanation are in the form of input features contributing to that prediction. For example, feature contributions in a text classification task have been used to show which word in the input data had the highest contribution in the context of the current output label.[9] Other examples of feature contribution explanations include pixels and super-

pixel weights in image classifiers[8,9] and attribute weights in tabular data prediction.[10] In our study, the XAI system also uses instance explanations to represent model reasoning about how it classifies news statements. Our system generates instance explanations in the form of confidence of veracity, attribute contribution, and related examples from training data.

Explainable user interfaces

The purpose of explainable interfaces is to present machine-learning results together with the machine-generated explanations for the end users. Examples of explainable interfaces have been implemented for recommendation systems[28] and personalized agents.[29] Research shows that the transparency of explainable interface algorithms can significantly increase user acceptance and reliance on machine recommendations.[30] For instance, Berkovsky et al[31] conducted a crowd-sourced study to examine the impact of recommendation-system strategies in terms of presentation, explanation, and priority on user's trust in a movie recommendation system. They measured user trust with nine key factors and showed a substantial difference among users based on their personality traits.

Along the same line, XAI research also studies users' mental models and performance in explainable interfaces with different explanations types and levels of detail.[14,28] For example, to study users' mental models and trust, Lim et al[15] designed and evaluated four different types of explanations in explainable context-aware systems. Their experiments indicate the importance for explanations to communicate *why* the system produces a given result in order for users to develop sound mental models. In another work, Kulesza et al[32] incorporated user feedback in an exploratory debugging interface to enable a two-way exchange of explanations between an end user and a personalized agent. In their study, user feedback from participants working with the explanatory debugging interface was roughly twice worthy compared to the feedback provided by the control participants in detecting model weaknesses. Another study of user feedback by Honeycutt et al[33] yielded evidence that soliciting interaction with or increasing attention to explanations to correct problems with the model caused users to become more skeptical of the model, which ultimately reduced trust and user perception of the model's accuracy. Other research of user trust and understanding of explanations suggest that users will underestimate a model's accuracy if explanations indicate the machine logic for classification does not match expected human logic.[34]

Similar results with domain experts have seen similar concerns when human operators do not understand machine logic, further motivating the need for explanation in critical domains. For example, a case study with computer network analysts found cybersecurity experts did not want to use algorithms they did not understand.[26] In this work, the addition of explanatory views for an anomaly detector was reported to increase user adoption in an operational analysis center. In other research, Krause et al[35] implemented an explainable interface for scientists to improve predictive models for detecting diabetes from medical records. In a case study, they showed how data scientists made use of an explainable interface to understand model behavior with localized inspection of the system. Controlled experimentation has also yielded evidence of differences in how domain experts and novices build and lose trust over time.[36] From their experiment with an entomological classifier, early observation of poor model outcomes caused more knowledgeable users to quickly lost trust in an intelligent system. Furthermore, when trust was lost early on, users were slow to increase their confidence in the system.

Other research of explainable intelligent systems has shown the importance of not only evaluating the explanations themselves but also studying how explanations may influence human thinking and decision-making. For instance, in research by Nourani et al,[37] a user study with interactive video review for activity recognition provided cautions about cognitive biases and overconfidence due to the addition of explanations. Such concerns demonstrate that it is not sufficient to simply include explanatory information without better understanding how the type of explanation might influence user behavior and their impressions of the model's capability. We also design and evaluate an explainable interface and evaluate user task performance and trust on the system with different types of explanations, but our core contribution is the controlled comparison of more nuanced variations in level of explanation.

News checking

False or fake news refers to media associated with propaganda, defamation, and material made to inflict emotional distress.[11,12] Fake news often has information that can be verified as false and is created with a dishonest intention.[38]

However, verifying whether a news statement is true is a difficult task. Recent events regarding web generated and propagated fake news have brought the issue to the attention of academics, and politicians, and citizens. The upsets from results in the 2016 US elections and the Brexit referendum have prompted calls for new regulations on web sites, advertisers, and social media companies. Starbird et al found that even journalists, who are trained to fact check, often unwittingly bring attention to rumors and false statements from Twitter.[13] While they later correct their reporting on the same channels, a rumor tends to gather more attention than its errata. Research has studied the characteristics of fake news in social media that lead to better understanding based on news content and social context.[39]

Previous research has demonstrated that AI and machine learning can potentially evaluate the veracity of news. Fact-checking systems can be classified into several areas such as spammer and bot detection, fake video, and image detection,[40] click bait detection,[41] rumor,[42] and truth discovery.[43] Fact checking has its own unique challenges for different data formats and purposes. For example, challenges and techniques in knowledge-based models (eg, knowledge graphs and open web sources) for truth discovery are far different from detecting fake video content. In data-based fact detection, Shu et al[44] introduced a method for detecting fake news articles using a variety of methods on the same corpus, such as linguistic inquiry and word count, in the context of supervised machine learning. Other detection schemes have addressed Twitter statements,[45] trust and labeling,[46,47] fake news-related images,[48] and how news is spread online.[13,49-51] Despite the difficulty of detecting fake news, research has shown lexical features can classify the overall truthfulness of statement.[52]

Our study utilizes an explainable news classification system. Our goal was not to build the most accurate fact detector, but to explore how different explanation designs affect how participants perceived and used recommendations. In other words, we prioritized designing a system that could partially explain itself over a method that could potentially produce better performance. Our approach is somewhat similar to[52] in that we also use PolitiFact data,[53] but we are providing an explainable system and studying how participants interact with it. Also relevant to our work is a study of agent-assisted news review by Mohseni et al.[54] This experiment studied a variety of explanation types together for an ensemble model to aid an open-ended review of entire news articles. With a variety of explanations available to users in the study, the results showed how simple forms of explanation helped users to identify poor aspects of a fact-checking model. Our current study builds on this work with further empirical study of simplified explanations but focuses more on amount of explanation and employs a more controlled experimental context with fact checking rather than news curation.

EXPLAINABLE FACT-CHECKING SYSTEM

The goal of our research was to study how different variations in machine explanation influence human performance and model understanding using an XAI system. To this end, we designed an experiment based on a fact-checking task and a corresponding XAI system. In this section, we describe the XAI system used for the experiment—an explainable fact checker to classify news statements as true or false. To meet the research goals for the experiment, we designed the XAI system's interface with explanation components we could reconfigure to create different experimental conditions.

News statement data and classifier

The data used in the study and to train the model were from PolitiFact, a fact-checking organization. Fact-checking news organizations offer independent assessments of claims made by politicians, often refuting their claims.[53] The interpretable text classifier in our XAI system was trained on a corpus with the size of 3500 samples from PolitiFact.com which has expert reviewed labeled news. On PolitiFact, each news statement is labeled as one of multiple classes, but we separate them into either more true (*true*, *half-true*, *mostly true*) or more false (*mostly-false*, *false*, *pants-on-fire*). This makes our classification task binary.

We note that the goal of this research was not to build the most accurate fact checker, but to study how different explanation designs affect the perception and utility of recommendations. Fact checking is a relatively new application of machine learning. In the political theatre, the epistemology of facts is not as rigid as they are in sciences, making logical axioms difficult to address. In terms of state-of-the-art accuracy, there are no central corpora of test data for comparing fact-checking agents. Approaches vary from analyzing sources (eg, URL links[55]) qualities of text for linguistic qualities,[56] and providing people with documentation for further investigation.[57]

Because the focus of this paper is on the human-subjects experiment and its findings, we limit the description of our system to an abbreviated summary, though we aim to provide enough details that a skilled machine-learning system designer could develop a similar system. For a more detailed description of the machine-learning aspects of the system, we refer the reader to the following demo paper[58] that describes the XFake architecture and technical details in full.

Our system's framework consisted of several models, and our final results for the AI Meter are ensembled from three models. All three models were given statements and their metadata, such as who the speaker, the context (eg, a debate), and what the statement targeted (eg, Obama). One model is the *MIMIC* framework that analyzes news metadata attributes and statements mapped to a GloVe word embedding[59] and a deep learning that then trains a student model consisting of 80 decision trees. The *ATTN* model uses a pretrained word2vec word embedding, convolutional neural network, and self-attention mechanism.[60] This is used to capture global relationships among words efficiently. The *PERT* framework looks at news statements using linguistic features, such as adjective ratio, verb ratio, sentiment, and scores. We also used a shallow model to mimic the deep learning model in a more interpretable fashion using an XGBoost classifier.[61]

The classifier output provides the overall confidence score (used for the AI Meter) and to model the attributes' salience (which we used for the attributes explanation type).[62] With the testing on validation set, we obtain 67.1%, 67.3%, and 53.2% accuracy, respectively, for MIMIC, ATTN, and PERT. We normalized weights according to the performance and fixed the coefficients to 0.36, 0.36, and 0.28 for the AI Meter and classification score. To provide feature contributions to explain the influence of specific *attributes*, for each prediction, we used a perturbation-based method that measures feature importance by looking at how much the classification score drops when a feature is removed from the input. Finally, our explanation method identified relevant *examples*; we used the most salient words as queries to the training corpus of example statements, then ranked them based on the highest number of shared metadata attributes.

Explanation interface

Changing how visualizations present, omit, and encode information often impacts how end users interpret their views.[63] We implemented our interface as a web application that provides the classification output and explanations for a given statement. As an XAI system, the interface also provides multiple types of explanatory information for the system's prediction. Figure 1 shows an overview of the interface with annotations for specific components. The *Statement* component shows an instance of input—the news statement and metadata of the *subject*, *context*, *speaker*, and *target*. The *Statement* shows the content of a stated fact or assertion about a Subject, in a particular Context, originating from a Speaker, by a particular Speaker, and sometimes directed at a Target. In deception, statements can provide linguistic hints or "tells" that both people and algorithms recognize as indicators of deception.[64] The *Subject* attribute includes categories related to the statement, such as "Education," "Crime," "Jobs," or "Poverty." The *Context* attribute includes

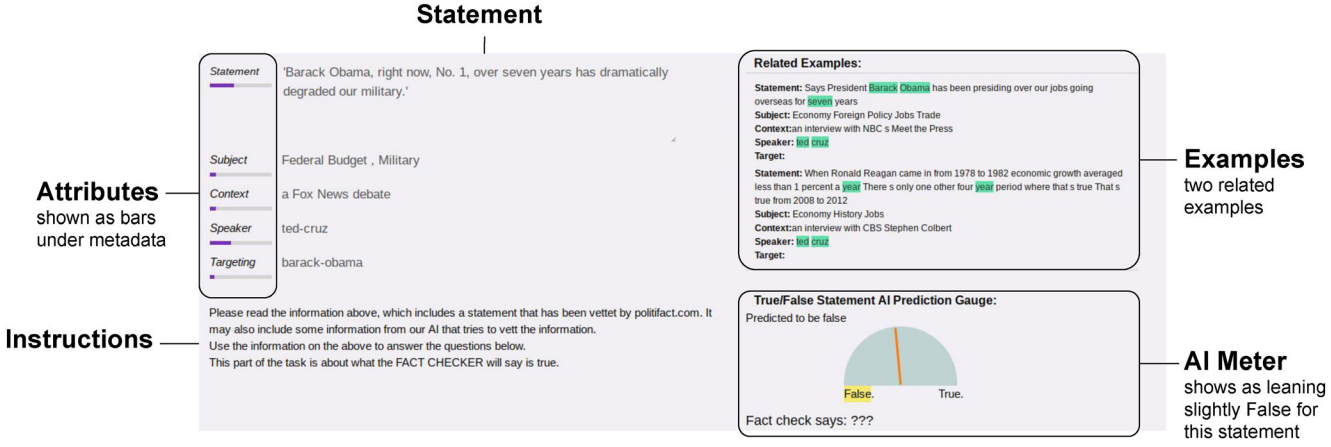

FIGURE 1 This image shows an example from the Fact Check phase under the AI Meter condition *On* and the Interface Component condition *Attributes + Examples*. All participants saw the *Statement*, which includes the news statement and other relevant metadata. The *AI Meter* component shows the overall guess via a dial. The *Attributes* shows how influential different statement metadata is for the given instance. The *Examples* component shows relevant alternative statements with the same classification

the location or event related to where the statement was made, such as "at a debate" or "on Fox News interview". The *Speaker* attribute includes the name of the person who made the statement, such as "Bernie Sanders" or "Hillary Clinton." The *Target* attribute, which may or may not be present for a given statement, refers most often to the person the statement is about (eg, as in when one politician is talking about another).

Different interface components provide different forms of explanation to help users understand the system's reasoning for making specific predictions. Overall, definitive guidelines for whether to show or hide an AI's confidence have not emerged.[53] Our thinking is that showing the confidence through an explanation will help participants have appropriate levels of trust.

The *AI Meter* interface component includes a dial representation that resembles PolitiFact's "Truth Meter." While many representations of uncertainty are possible,[65] we decided to make the AI Meter similar to the Politifact "Truth Meter" in order to provide a simple and familiar representation of our model's prediction of the truthfulness of political statements. Politifact's "Truth Meter" is represented as a dial and explicit levels of text that show how truthful they rate a political statement.[66] Generally, statements are in political arenas are not completely factual. Intentionally dishonest statements weave in falsehoods into otherwise honest content.[64] Likewise, the Politifact Truth Meter rates the honesty of statement across six levels from "pants on fire" (the most untrue) to "true" (the least untrue). Our AI Meter includes a dial that is mapped to the overall classification value from the XAI output. In this way, the AI Meter provides an explanation of model confidence. The AI Meter's indicator needle points to the left for "false," to the right for "true," and straight up when "unsure." This component also highlights either the "False" or "True" labels to emphasize the overall binary recommendation.

To provide further explanation, the *Attributes* interface component shows the relative importance or salience of the different attributes in news statements in terms of how they influenced our XAI model. For example, one particular news statement may have the *speaker* and *statement* attributes be the most influential, whereas a different news statement might have the *targeting* attribute as most influential. By "influential," we mean that our modeling has found these to impact veracity the most. Various aspects of context are captured by the attribute, for example, "Speaker." Attributes are tied to metadata, which can be important when determining veracity. Through a series of experiments, Blair et al found that reviewing contextual information significantly improved participants' ability to detect lies.[67] The Attributes explanation help provide information about which metadata is most important for determining the veracity of a statement. In a recent survey of rumor detection in social media, Zubiaga et al found that around half of automated solutions incorporate metadata attributes.[68] Because attributes empirically improve performance in prediction systems, we reasoned that showing the relative importance of attributes provides a form of explanation.

Finally, an additional form of explanation is given by the *Examples* interface component, which uses keywords to present additional related statements from the PolitiFact training set data with a similar classification output as the statement instance. Providing this information, we expect, would help end users get a sense of the range dialogue around similar subjects. Our systems show two related example statements with relevant keywords highlighted.

Kurtz et al, through their experiments and discussion, find that examples are essential for learning.[69] Examples serve as a baseline for the range of possible and potentially analogous, situations. Cognitively, this can teach people to solve problems by drawing on their recollection of similar situations their conclusions. Evaluating the veracity of news statements may benefit from an example-based reasoning approach. As people are exposed to news, priming effects can anchor[70] their expectations. As Wall et al explain, bias can be seen as a way to model understanding.[71] Valdex et al shows that, when presented with a sequence of example visualizations, a person's evaluation baseline changes.[70] Our Example component also displays the political speaker, which involves social information. Social information, such as who is sharing information, impacts how well they are trusted.[72] Thus, we reasoned that showing examples provide participants with additional context that would impact their veracity estimations and understanding of the classification system.

The interface can be configured to provide or exclude different explanation components. We describe the experimental conditions and configurations in the following section. For the purposes of the experiment, the interface presented to participants also included instructions, questions, and buttons that guided participants through tasks as we collect their responses.

EXPERIMENT

We conducted a controlled experiment to evaluate how different amounts of explanation provided by an XAI system would affect user task performance and model understanding.

Goals and hypotheses

The experiment was designed to address two primary research questions: (1) "How does the amount of explanation information affect human performance in assessing the veracity of news statements?" (2) "How does the amount of explanation information affect user understanding of the XAI models?"

We hypothesized that providing too much information as part of the system explanation could be overwhelming and therefore reduce its utility, which we would expect to negatively affect human performance and decision-making. On the other extreme, we predicted having some explanatory information would be better than no explanation. We hypothesized that a moderate amount of explanation is important for achieving the benefits of XAI, and understanding the appropriate balance of useful information is important for designing explanation interfaces. Thus, we designed an experiment to test different types and levels of explanation detail in order to evaluate the effects on participant performance and model understanding.

To test these hypotheses, we needed a challenging human analysis task that could potentially benefit from the assistance of an intelligent system. We chose the task of news-statement fact checking as the basis for the experiment due to its overall difficulty and relevance in society that require human judgment. In addition to our focus on the above hypotheses, we also wanted to learn more about human review and decision-making processes used when evaluating the truthfulness of news statements. A better understanding of how people assess and make sense of news statements is important for future improvements to XAI systems and explanation design. For this reason, we solicited qualitative feedback aimed to identify the issues involved and strategies that may be used in the process of detecting fake news.

Experimental design

To address our research goals, we designed an experiment that asked participants to review news statements with different interface configurations that varied explanation types. Refer to Figure 1 for the corresponding explanation components. To test the most simplistic form of explanation, we controlled for the presence or absence (ie, *on* or *off*) of the *AI Meter*, which showed the model confidence in the predicted classification.

We also varied additional levels of explanation detail based on the inclusion of attribute weights (shown visually by colored bars for each attribute; see upper left of Figure 1) and supporting examples (shown by similarly classified statements; see upper right of Figure 1). Using these explanation components, the experiment controlled three levels of supplemental explanation via *Explanation Type*: *None*, *Attributes*, and *Attributes + Examples*.

Thus, the experiment followed 2×3 between-subjects design with two levels of the AI meter factor and three levels of supplemental explanation vai Explanation Type. In our between-subjects design, each participant experienced only one version of the system. This provided consistency between paired phases, where participants first gained experience reviewing the explanations (Fact Check), then used this experience to predict how they XAI model would assess news statements (Prediction-Guess). Our design allows us to configure different levels of explanatory information and compare how it impacted participant performance and understanding across conditions.

Tasks and measures

To collect measures for human new-verification performance and understanding of the model, we designed the experiment to have two primary phases: Fact Check and Prediction-Guess. Figure 2 shows an overview of the study procedure and how the tasks were influenced by the experimental conditions. While there are two main phases, we also use questionnaires that ask participants to answer questions about their demographics, political leanings, news consumption, and their experiences using our XAI systems.

In the *Fact Check* phase, participants used the XAI system to review news statements (one at a time) along with the corresponding output and explanations per the assigned experimental condition. Before the Fact Check phase, participants were shown tutorial slides that highlighted the interface components (AI Meter, attributes, examples) that differed based on the experimental condition. The Fact Check phase served as a short training session for using the XAI system. Participants were tasked with assessing whether the news statement was true or false; they were free to decide whether or not to agree with the AI classification. After they evaluated a statement, the system showed the correct answer before allowing them to proceed to the next news statement. Participants evaluated 16 news statements in this

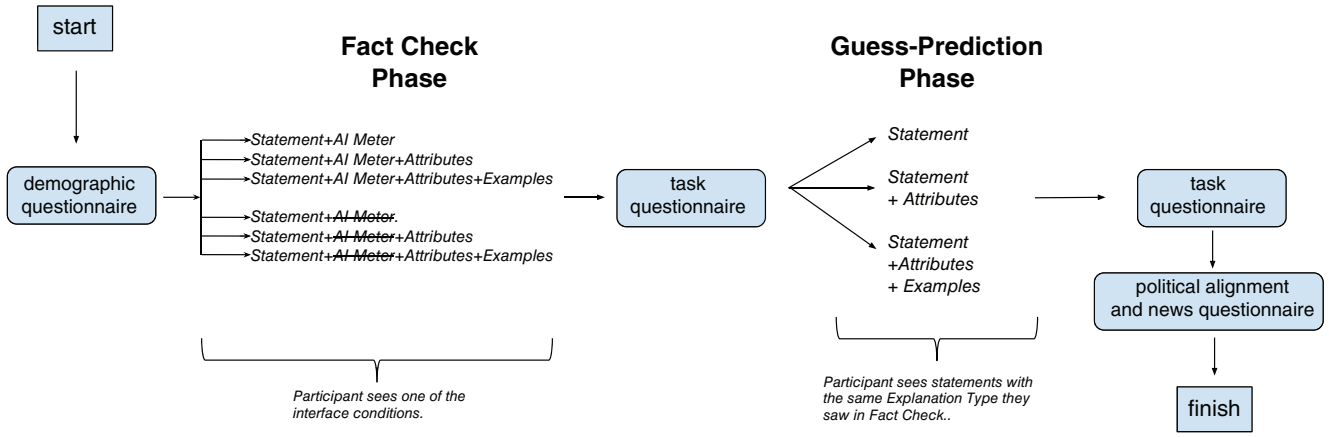

FIGURE 2 A summary of the experiment's procedure as experienced by participants

phase. As dependent variables, we recorded correctness and response time as measures of human performance. After assessing all statements in this phase, participants also provided subjective feedback about their experience via a questionnaire. We note that, in the Fact Check phase, we provided PolitiFact's answer after participants rated a news statement (eg, replacing the "???" at the bottom right of Figure 1). This was true across conditions. We do not include the time between reviewing these results in our response time measure.

Following the Fact Check phase, participants then moved to the *Prediction-Guess* phase. In this phase, participants again viewed 16 news statements, except their task was now to predict the classification output of the XAI system. In other words, instead of predicting the statement veracity, the goal was to predict how the model would assess the statement. This task was done as a way to evaluate participant understanding of the model based on their ability to infer its output for a new set of input. Trials in this phase did not provide the AI classification output, and no feedback was given about whether their predictions were correct. Participants provided their answers using an interactive version of the AI meter that allowed them to adjust the position of the indicator needle to the predicted location, which provided data for *true* and *false* predictions and to indicate the level of system confidence. The *supplemental explanation* factor also influenced how much explanation participants experienced in the Prediction-Guess phase. As in the previous phase, participants evaluated 16 new statements in this phase, and time and correctness (aligned with the model's guess) were again recorded for this prediction task. For this phase, we did not provide an answer after participants estimated the AI Meter.

We used questionnaires between main phases to collect information about the participants and their qualitative experiences. Before starting the Fact Check phase, participants answered questions about their demographics. After each main phase, participants answered task questionnaires that asked participants about their experience. For example, we asked them what their general strategy was during their tasks. Finally, we asked participants questions about their news reading habits and political alignment.

Experiment Data Corpus

The experiment used the XAI system results and explanations for news-statement data. The experiment used the same PolitiFact data[53] as previously described. However, to avoid confounds and maintain experimental control, we controlled the data composition to ensure all participants reviewed equal numbers of true and false statements with the XAI system and that the system performed at a fixed level of classification accuracy. Each participant completed the tasks with 32 news statements drawn from a pool of 100 news statements that had been previously evaluated by the XAI system and cached locally. In our method, the specific statements varied among participants, but all participants completed the study with the same composition of true positives (37.5%), true negatives (37.5%), false positives (12.5%), and false negatives (12.5%) from the total corpus. In order to evaluate the explanation system, it was important that participants observed both correct and incorrect classifications. For the purpose of our experiment, we controlled the amount of accuracy experienced by all participants to (1) remove noise in our data and (2) to ensure participants had

reason to at least partially trust the system. Early pilot testing showed that participants seeing 50% correct (equal portions of true positives and true negatives) and 50% incorrect (equal portions of false positives and false negatives) accuracy mistrusted the model. Because of this, we adjusted this mixture to 75% correct and 25% incorrect classifications.

Participants

We conducted our experiment online using Amazon Mechanical Turk.[73] We paid each participant $5 USD. Based on the time it took to complete the study, participants made more than minimum wage. We set this amount based on pilot studies, which took an average of around 30 minutes to complete. Through Mechanical Turk configuration, participants were required to have completed 100 prior HITs, have at least 90% acceptance rate, and be located in the United States. We recruited a total of 190 participants. For quality assurance, in addition to the constraints on participants, we implemented an attention check problem during the Fact Check phase. We paid all participants, including those that failed the attention check.

After filtering for attention check failures, 147 participants were included for data analysis. Of those, 101 identified as male and 46 as female. Their ages ranged from 20 to 72, with a median of 32 years. Participants had various levels of education, with 22 reporting a High School or equivalent education, 36 having some college, 67 with a college degree, and 22 with a Masters or higher degree. Polls have shown more education correlates with more liberal political alignment.[74] In terms of political alignment, 27 participants considered themselves moderates, leaning neither left nor right, 90 were liberal, and 30 were conservative. Of our 147 participants, 120 said they paid attention to the 2016 US presidential election, which may also indicate familiarity with the concept of fake news. The vast majority viewed their news from news websites and social media (112), rather than television (28), radio (4), or print (3). Overall, this shows our participants were fairly sampled from a range of political leanings, education, and news source consumption.

RESULTS

We report findings based on both quantitative and qualitative results. Quantitative measures met the assumptions for parametric testing (ie, normality and homogeneity of variance), so we report statistical analyses using two-way independent factorial analysis of variance (ANOVA) tests in accordance to the experimental design to assess evidence of the effects due to type of supplemental explanation and the presence of the AI Meter. For brevity, we report test statistics only for significant effects. To help summarize participants' understanding and strategies, we qualitatively coded the results from questionnaires and organized the results based on their volume and relevance to our research goals.

Fact Check results

In the Fact Check phase, participants rated news statements as true or false. We measured the accuracy of participants as the total number of answers right divided by the number of news statements (see Figure 3 and Table 1). A majority of participants performed well over random (ie, 0.5), with all mean accuracy well above 0.5. Still, the results indicate that the task was difficult, which was not surprising given that the news statements were short and fact checking generally involves consulting outside information. A two-way ANOVA found no significant difference in accuracy among conditions for AI Meter and Explanation Type, and no interaction effect was detected. Thus, this result does not support our hypothesis that added explanation supports improved human task performance; however, this is logical due to the difficulty of the task. One participant, without the AI Meter and in the None Explanation Type condition, mentioned in the general strategy:

> Without knowing what actually happened, it's impossible to know if a statement related to it is true.

Based on our review of qualitative responses from participants, they were clearly engaging in the task. For example, a participant who had the AI Meter said:

> Where I didn't know [the answer], I took the AI's guess ... into consideration ... to help me decide.

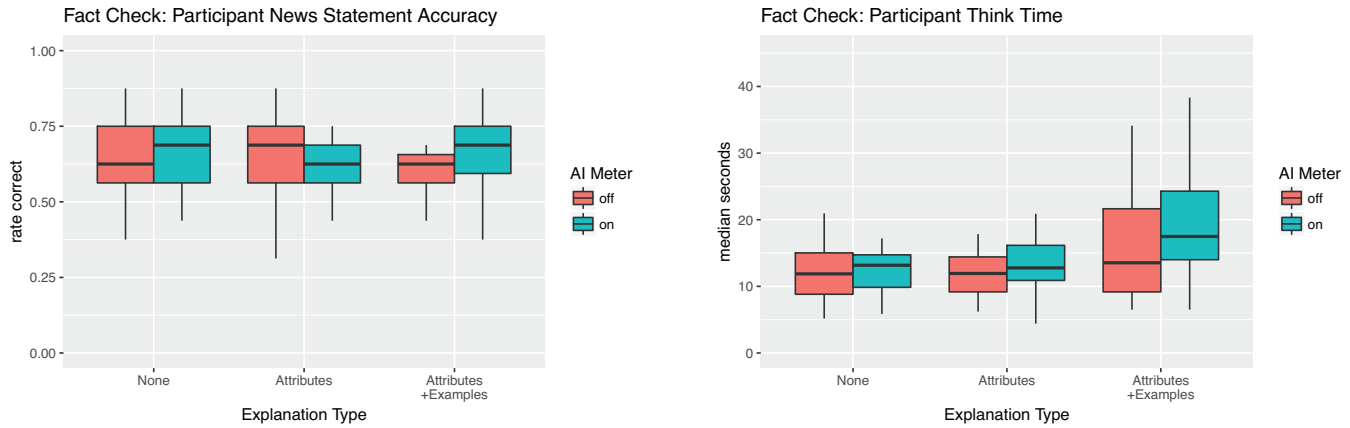

FIGURE 3 Left shows boxplot of accuracy of participants in Fact Check phase, measured as rate correct against PolitiFact. Higher values indicate better performance. No significant differences were detected across conditions. The right shows median participant response time for each statement in the Fact Check phase. The *Attributes + Examples* condition was significantly higher than both the *None* and *Attributes* conditions

TABLE 1 Performance summary for Fact Check task showing the mean (M) and SE for accuracy and time across conditions

(a) Condition		Accuracy		Time	
Explanation	**AI Meter**	***M***	**SE**	***M***	**SE**
None	Off	0.64	0.14	12.6	5.59
Attributes	Off	0.65	0.15	13.1	6.22
Attr + Examples	Off	0.60	0.10	16.7	8.45
None	On	0.67	0.11	12.6	4.11
Attributes	On	0.60	0.15	14.4	6.47
Attr + Examples	On	0.65	0.15	20.5	9.87

(b)	Accuracy		Time	
Condition	***M***	**SE**	***M***	**SE**
AI Meter On	0.64	0.14	16.1	8.08
AI Meter Off	0.62	0.14	14.0	6.96
Explanation: None	0.65	0.13	12.6	4.90
Explanation: Attributes	0.62	0.15	13.7	6.31
Explanation: Attr + Examples	0.63	0.13	18.7	9.35

Note: In (a), we show a six-way split for each Explanation × AI Meter condition. In (b), we show a two-way split between the AI Meter as well as a three-way split among each explanation.

The results of a two-way ANOVA for the rate of participant agreement with the AI Meter provides evidence for this behavior. We measured agreement as the number of statements when the participant's and AI Meter's rating were the same. Participants with the AI Meter ($M = 0.73$) had a significantly higher rate of agreement than those without ($M = 0.60$), with $F(1, 141) = 31.12$ and $P < .001$ No effect was detected for Explanation Type, and no interaction effect was detected.

This indicates that participants were paying attention to the AI Meter. In this way, having an AI Meter impacted how participants performed the task. Our review of qualitative data shows examples of participants using the AI Meter as a tie breaker.

We also assessed the time it took participants to perform each rating of a news statement (see Figure 3 and Table 1). To account for response variability of online studies, we use each participant's median time for analysis. We removed one outlier that had a time of more than 75 seconds for this measure. A two-way ANOVA test found a significant

difference for Explanation Type with $F(2, 141) = 10.4$ and $P < .001$, and but no effect for AI Meter with $P = .139$. The interaction was also not significant. A post hoc Tukey's honest significant difference (HSD) for Explanation Type found the *Attributes + Example* condition took significantly longer than both *None* ($P < .001$) and *Attributes* ($P < .01$). We believe the reason for this is that these Explanation Types engaged participants more, taking their attention for news statements they were unsure of, but this engagement required more time to review and interpret the output.

Our results for the Fact Check phase shows that the amount of information offered to users can significantly impact the time it takes to perform tasks. Participants in the conditions with the most supplemental information took more time to complete the same tasks, compared to *None* and *Attributes*.

Prediction-Guess results

For the Prediction-Guess phase, we measured the participant's estimation error and time while they guessed the prediction using an interactive AI Meter (see Figure 4 and Table 2). In other words, instead of seeing an AI Meter with a recommendation, participants set the position of the dial with a slider. For each news statement in the Prediction-Guess phase, we measured error as the distance between the participant-set value and the AI Meter's recommendation. For example, if the AI Meter's values were fully false (−0.5) and the participant selected a value of fully true (0.5), the distance would be 1.0. We refer to "estimation error" as the mean of these distances between participant estimations and the hidden AI Meter.

A two-way ANOVA found significant differences in estimation error for Explanation Type (None $M = 0.25$, Attributes $M = 0.25$, and Attr + Examples $M = 0.21$) with $F(2, 141) = 3.69$ and $P = .027$, and for AI Meter (on $M = 0.20$ and off $M = 0.27$) with $F(1, 141) = 32.0$ and $P < .001$. No interaction effect was detected between the two factors. Following the significant main effect of Explanation Type, we compared types of explanation when the AI Meter was present. A post hoc Tukey HSD found estimation error was significantly worse (ie, less accurate) for *None* ($M = 0.22$) than *Attributes + Examples* ($M = 0.17$) with $P = .02$. This means participants with the most supplemental explanation information were more accurate in the Prediction-Guess phase. Likewise, the Fact Check accuracy for participants with the AI Meter were slightly higher ($M = 0.64$) compared to those without it ($M = 0.62$), but this result was not found to be statistically significantly different.

We also considered response time for tasks in the Prediction-Guess phase (see Figure 4 and Table 2). Due to participant variability in the online procedure, we analyzed each participant's median time used to set the AI Meter for each news statement. A two-way ANOVA failed to detect effects due to AI Meter, and no interaction effect was found, but the test nearly finds significant differences in time for Explanation Type with $F(2, 141) = 2.80$ and $P = .064$. A post hoc Tukey HSD for Explanation Type found the *None* condition took nearly significantly less time than *Attributes + Examples* ($P = .063$). While, in our analysis of the Fact Check phase, we found a significant difference between *None* and *Examples + Attributes*, this result is in the same direction, but nearly significant. Again, participants with the Explanation Type with the most supplemental information took longer to perform their tasks.

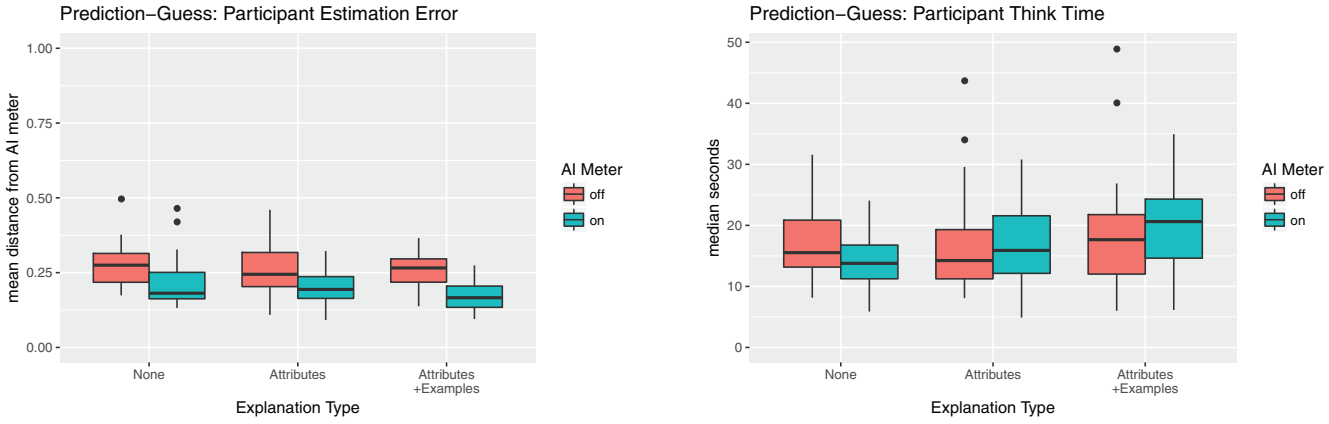

FIGURE 4 Left shows the estimation error in the Prediction-Guess phase. Here, lower values are better. Participants had significantly higher performance with the AI Meter. Participants with the *Attributes + Examples* Interface Type were more accurate in guessing what the AI Meter value would be. The right shows median response times for Prediction-Guess trials. The Explanation Type shows a near significant difference, shorter for None and longer for Examples + Attributes

TABLE 2 Performance summary for Prediction-Guess task showing the mean (*M*) and SE for error and time across conditions

(a) Condition		Error		Time	
Explanation	**AI Meter**	***M***	**SE**	***M***	**SE**
None	Off	0.28	0.08	17.4	6.24
Attributes	Off	0.26	0.09	17.1	8.57
Attrs + Examples	Off	0.26	0.06	18.7	9.85
None	On	0.22	0.09	14.2	4.72
Attributes	On	0.20	0.06	16.3	6.33
Attrs + Examples	On	0.17	0.05	19.6	7.07

(b)	Error		Time	
Condition	***M***	**SE**	***M***	**SE**
AI Meter On	0.20	0.07	16.8	6.51
AI Meter Off	0.27	0.08	17.7	8.22
Explanation: None	0.25	0.08	15.8	5.74
Explanation: Attributes	0.23	0.08	16.7	7.50
Explanation: Attr + Examples	0.21	0.07	19.2	8.39

Note: In (a), we show a six-way split for each Explanation × AI Meter condition. In (b), we show a two-way split between the AI Meter as well as a three-way split among each explanation.

When interpreted along with the error estimation results, the time results for the Guess-Prediction task suggest a trade-off between ease of fast comprehension and level of understanding. These prediction results partially support our hypothesis that additional explanation information can improve understanding of the model. The error estimation results did not support our hypothesis that the highest level of explanation detail would be detrimental to understanding, but the time results did indicate that the highest level of detail did take longer for participants to process.

Qualitative results

The questionnaire following the Fact Check phase asked participants to respond to open-ended questions about their overall strategy. We reviewed all of their responses and performed a bottom-up process to distill descriptions of the strategies participants used. One strategy involved *feeling and intuition*. These participants reported using a nonspecific combination of background knowledge and what they thought might be important. They emphasized how true statements "felt" and whether they "made sense." For example, one participant said:

> I didn't have one strategy, I just tried to think about what made sense overall....

Another strategy was *simple heuristics*. For example, participants reported basing their ratings on prior knowledge about the speaker. Others reported that "blanket statements" were more often false.

> If the speaker was Donald Trump, I generally leaned towards it being false.

> If I didn't have a direct knowledge of the fact then I relied upon my beliefs if the reliability and bias of the speaker.

> ...Blanket statements were often false at the beginning, so I went with that.

Along with reports of strategies, participants often mentioned the issue of trust with the AI Meter and Explanation Types. The trust in the AI was mixed.

> I did use the AI and it did help me since it was more right than wrong many times I grew to trust it.

> The AI was right most of the time so I trusted it.

> I felt the AI gauge was not reliable so did not use it for most of my decisions.

Participants in conditions with the AI Meter and Explanation Type used them as *tie breakers*. In general, participants relied on their intuition and prior knowledge about news statements. However, the AI Meter and Explanation Type would influence them in cases where they felt unsure.

> I mainly used [the AI Meter] if I was unsure. Kind of like a tie breaker.

> I used [the AI Meter] as a check on my first impression, depending on how strong I felt and how strong the AI's opinion was, I took it all into consideration for my final decision.

We also analyzed all responses from participants about their strategies for the Guess-Prediction task. Overall, participants' strategies were similar to those used in the Fact Check phase. This was due in part to participants assessing the accuracy of news statements rather than attempting to estimate the AI Meter directly.

> I just went by if it sounded true or not and if the AI would think if it was true or not.

> I used my beliefs and desire for truth and imposed them on what I thought the AI would guess.

In terms of the task of estimating the AI Meter, participants reported a range of strategies and thought processes. Simple strategies included paying attention to the speaker. More nuanced approaches altered participant estimation based on the content of the statements. For example, participants reported moving the dial on the AI Meter based on how "factual" or "strong" statements were.

> It seemed like the AI would be more confident on a cut and dry matter of fact as opposed to something containing any sort of opinion.

> I had a hard time pinning down exactly what I thought the AI would use, but stronger statements seemed to have higher influence on it, so I used that as my basis.

> If it was true and bizarre, then I felt the AI would be more likely to rate it as false even if it's true.

Another participant tried to emulate how they perceived the AI would interpret keywords and phrases.

> I tried to interpret the information in a way that I feel like an AI would understand, using certain keywords and how the information was phrased. I think this allowed me to somewhat accurately predict what the AI would guess.

Asking questions about their strategies soon after they completed each phase provides a unique opportunity to understand how users might perceive and work with XAI systems. Overall, the qualitative results show that participants were often engaged in their tasks and thought deeply about their decision-making processes.

DISCUSSION

Our primary research goal was to understand "how the amount of explanatory information provided by an XAI system impacted human performance" in a data review and decision-making task. Overall, the results show that fact-checking news statements—with the restrictions of not consulting outside information—to be very difficult.

Fact Check, Prediction-Guess, and mental model results

The results of the *Fact Check* task did not show significant differences in accuracy due to the amount of explanation information. However, participants with additional information spent significantly more time during the *Fact Check* and the *Prediction-Guess* tasks. This indicates that there is an attention cost when asking users to perform tasks with supplemental information about AI.

For the *Fact Check* phase, participants described using their own intuition first and relying more on the AI Meter and Explanation Types when they were less sure. Participants described their "intuition," "feelings," and "hunches" as an important first step. One said it served, "as a check on my first impression." They also employed their own heuristics about the speaker and factual nature of claims.

The Prediction-Guess phase more closely addresses our second question of "how the amount of explanation information affect user understanding." Both the presence of an *AI Meter* (during the Fact Check phase) and *Explanation Type* (present during both phases) impacted participant performance in terms of estimation error. Participants were worse at predicting the AI output if they previously completed the first task without the AI Meter, which suggests that participants were learning about model outputs for given inputs. This finding provides general support for the fundamental notion that XAI can facilitate improved understanding of intelligent systems. Perhaps more interestingly, participants made better predictions of model outcomes in explanation conditions that offered the most supplemental information. Like the *Fact-Checking* task, they took significantly more time on the *Prediction-Guess* task in the conditions of *Explanation Type* that provided the most supplemental information.

In terms of addressing mental models, participants mentally sidestepped the task of estimating the AI Meter by assuming the AI would be correct. In other circumstances, they attempted to emulate the thought processes of the AI by looking for potential keywords, the distribution of attribute scores, and whether the statement seemed more factual or opinionated. We expect that users interacting with other XAI systems would develop similar ad hoc strategies for managing their own performance on tasks and the performance of the AI they are working with. Users bring their own experiences and external information with them for data judgment and decision-making tasks. When designing XAI systems, it is important to take into consideration that people will likely develop their own ideas about how the system works. Our results demonstrate that different users may adopt different strategies and interpretations for a system—even when using the same interface and same models.

Trade-offs in XAI interface design

We see a fundamental trade-off in how people design interfaces for XAI systems: more modalities may provide better understanding at the cost of attention and time. In our study, more information improved participant accuracy in the Prediction-Guess phase, the benefits came at the cost of time and attention. People who design XAI systems should be aware of this trade-off when designing for users. Having additional information in XAI systems can increase the overall performance of users, but may introduce a trade-off of attention and time needed to understand them.

Like many AI systems, ours was composed of subsystems that combine to create a final score used to make predictions. In our case, the visualizations corresponded to different components of our ensemble. The attribute visualization corresponded to the MIMIC's interpretation of salient information, the supporting examples corresponded to similar phrases and attributes, and the AI Meter shows the composite score. Our premise was that when visualizations provide visual and perceptual aids that change in proportion with the context of ML that they necessarily provide an opportunity for its users better understand what the system will predict.

We note here that the presence of the AI Meter provides the most improvement and, given the Prediction-Guess task, the best explanation. There may be several reasons that this is the case. The AI Meter is one of the most simple representations that is directly tied to how our XAI system models its understanding. This could represent an easier to remember and understand visual signal that more closely aligns with the reality of fact-checking as a somewhat subjective task. For example, more fact-based and noncontroversial statements may be easier to predict and the meter serves as a less binary representation of true and false. Another possibility is that practice in the Fact Check phase may have sensitized participants to be more conservative in their guesses.

While this paper presents one approach to developing explanations through a combination of components, future work might test a myriad of potential modalities. There is no end to the number of potential visual representations of state and system-internals that could impact the utility of an XAI system. For example, researchers might study

providing additional examples from external systems, ways to debug the system itself, and distribution of veracity based on the historical record of the speaker.

We also observe that the complexity of underlying models for machine learning may have a higher upper-bound on overall complexity, and therefore unintelligibility, compared to the visual features presented by interfaces that are naturally limited by human perception. The "more and more complex model architectures" associated with Deep Learning models will eventually make their perceptibility "invisible under the weight of layers of the model".[75] In contrast, heuristics generally have a low complexity and do not require complex interpretations. The only apparent solution to increase understanding is to simplify machine-learning models, which would cost accuracy and utility, or to design interfaces that visualize simplified correlates of model internals. This emphasizes the importance of inventing novel understanding and interfaces that can correlate to machine-learning internals.

Another potential trade-off for complexity depends on the context for the intended use case and who will be using the system itself. In our study, users conducted a verification task that allowed them to form their impressions of the system over the duration of the first phase of the study—providing a relatively limited and constrained experience with model capabilities. As study volunteers, participants likely had limited pressure in the task based on lack of consequences for incorrect decisions. In contrast, system usage in operational contexts would provide a longer period of usage (and likely multiple periods over time) to allow users' mental models to evolve with continued experience. In addition, the study's usage context represents limited user investment in the task compared to potential users of such systems including journalists, intelligence analysts, or political strategists. More involved participants with may be more motivated and trained to better work with particular XAI systems. Also related to user contexts, additional explanation approaches for explainable fact checkers could leverage differences in information needs based on user knowledge or expertise. Familiarity with the task of fact checking might lead to ideas for better interfaces. Prior work has found that with systems used to do deep dives, that people are willing to wait longer and have more attention over time.[76]

Future research could also support different types of user tasks for information verification and advancement of fact-checking models. We studied an XAI system by asking participants to estimate its predictions. Our system design summarizes feature and model attribution for specific input-output pairs rather than aiming to communicate a direct representations of AI internals. This contrasts other XAI systems for debugging finding problems in specific AI internal components.[32] Furthermore, participants in our study were not ML experts and were not offered detailed information about the creation and design of the AI's underlying models. Despite this lack of direct representation, they were more accurate in the Prediction-Guess phase as detail increased, thus demonstrating how instance-level explanations can help users to develop an intuition for understanding computational models through application-level experiences.

XAI effectiveness through empirical evidence of prediction

One of the basic goals of interaction design is to convey accurate mental models of systems to help align understanding of interaction possibilities with actual system operation.[77] However, given the "more and more complex model architectures",[75] it can be challenging (or impossible) to create explanations that faithfully represent model logic while still being simple enough for human understanding. Support for application-level human judgments with machine assistance can take advantage of designs that provide supplemental information that allow end users to naturally improve their understanding through continued observations of system operation. By this approach, rather than aiming for novelty in creating representations that entirely model complex internal representations, our research contributes an iterative and component-based design method that prioritizes high-level feature information to help users maintain the connection between represented information and model functionality.

This method also facilitates evaluation of a wide variety of explanation designs. In our research, we demonstrate the value of a broad approach that generates several new visual explanation representations and empirically tests them to evaluate effectiveness. This approach enables user testing in a fashion similar to a unit testing approach to further advance knowledge of a broad set of explanation strategies and representations. By taking advantage of the "user prediction" approach to evaluating model understanding by tasking users with predicting future AI guesses for new inputs, as our Prediction-Guess does, we can study model understanding empirically without regarding whether the participants' mental models match an AI's model directly.

With an empirical and pragmatic basis for evaluating understanding, new research has a tremendous breadth of possible visual representations and explanation types. The designers of these systems may have specific hypotheses about how they should be used in practice. However, as the complexity of AI models increase, it becomes less likely that

understanding its inner mechanics will be possible or have practical utility for application-level end users. Our participants reported ad hoc and holistic decision-making for prediction, that they "took it all into consideration" to predict. From an evolutionary perspective, this approach follows parallel cognitive mechanisms in Theory of Mind.[78] As humans develop skills to understand and attribute mental states to others, for example, by looking at what they do and relating it to their own experiences, they engage in Theory of Mind social cognition. The Simulation Theory posits that people accomplish this by using a combination of internal first-person thinking that is adjusted during real-world experiences. In other words, the qualitative results of our study may suggest that users of complex AI systems in other contexts will inevitably develop personalized heuristic-generating practices that mitigate the cognitive cost of understanding complex models. For future XAI research, the results suggest that the reception of user interfaces will take on a life of their own when combined with the social and intellectual practices of their users.

CONCLUSION

The presented research provides novel empirical data of how different types and amounts of explanatory information affects user ability to utilize explanations to understand system behavior and improve task performance. In a controlled experiment, participants were tasked with using an explainable fact-checking system to assess news statements and to predict the output of the AI. The study's results indicate a clear tradeoff between speed and accuracy due to added explanation details. On one hand, the results provide strong evidence of the value of additional explanation information for intelligent systems, as additional information did help users to better understand and predict system behavior. On the other hand, it took more time to review and interpret the results in order to achieve the benefits.

This raises additional questions about whether such a tradeoff could be avoided with alternative types of explanation. Explanation is difficult, and we do not claim that our studied XAI system follows an optimal design. It could be possible to design an explanation interface that effectively communicates a perfect amount of information to facilitate improved understanding without causing the penalty of additional human processing time. However, achieving an optimal design would likely require prior knowledge of the users' level of expertise, the level of explanation detail that would be most useful for one particular task, and the best explanation format for both the model and the user task. Such complexity and possibilities for design demonstrate the need for future research on the topic of XAI and human-interpretable systems.

ACKNOWLEDGMENTS

The 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.

DATA AVAILABILITY STATEMENT

This study was run from a model based on publicly available data from Politifact's API for research purposes, but is not approved for redistribution. The corresponding author can provide example data presented to participants on request, but not their answers without additional IRB approval.

ORCID

Rhema Linder https://orcid.org/0000-0003-4720-6818
Sina Mohseni https://orcid.org/0000-0003-2747-2377
Fan Yang https://orcid.org/0000-0003-3442-754X
Shiva K. Pentyala https://orcid.org/0000-0002-2985-9113
Eric D. Ragan https://orcid.org/0000-0002-7192-3457

REFERENCES

1. Höök K. Steps to take before intelligent user interfaces become real. *Interact Comput*. 2000;12(4):409-426.
2. Bellotti V, Edwards K. Intelligibility and accountability: human considerations in context-aware systems. *Human Comput Interact*. 2001; 16(2–4):193-212.
3. Gunning D. *Explainable Artificial Intelligence (XAI)*. Arlington, Virginia: Defense Advanced Research Projects Agency (DARPA); 2017.

4. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. *ACM ComputSurv (CSUR)*. 2018;51(5):93.
5. Lakkaraju H, Bach SH, Leskovec J. *Interpretable Decision Sets: A Joint Framework for Description and Prediction: SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM; 2016;22:1675-1684.
6. Rosenthal S, Selvaraj SP, Veloso MM. Verbalization: Narration of Autonomous Robot Experience; 2016: 862–868.
7. Zeiler MD, Fergus R. *Visualizing and Understanding Convolutional Networks*, European Conference on Computer Vision. Switzerland: Springer International; 2014:818-833.
8. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv Preprint arXiv:1312.6034*; 2013.
9. Ribeiro MT, Singh S, Guestrin C. *Why Should I Trust You? Explaining the Predictions of any classifier*, SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2016:1135-1144.
10. Ribeiro MT, Singh S, Guestrin C. *Anchors: High-Precision Model-Agnostic Explanations*; Association for the Advancement of Artificial Intelligence, California: AAAI Press; 2018;32:1.
11. Klein D, Wueller J. Fake news: A legal perspective. *Australasian Policing*. 2018;10(2):11-15.
12. Kucharski A. Post-truth: Study epidemiology of fake news. *Nature*. 2016;540(7634):525.
13. Starbird K, Dailey D, Mohamed O, Lee G, Spiro ES. *Engage Early, Correct More: How Journalists Participate in False Rumors Online during Crisis Events*. New York: ACM; 2018:105.
14. Stumpf S, Rajaram V, Li L, et al. Interacting meaningfully with machine learning systems: three experiments. *Int J Human Comput Studies*. 2009;67(8):639-662.
15. Lim BY, Dey AK, Avrahami D. *Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems*. New York: ACM; 2009:2119-2128.
16. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. *arXiv Preprint arXiv:1702.08608*; 2017.
17. Ross AS, Hughes MC, Doshi-Velez F. *Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations*. California: IJCAI; 2017:2662-2670.
18. Zhang Q, Wang W, Zhu SC. Examining cnn representations with respect to dataset bias. *arXiv Preprint arXiv:1710.10577*; 2017.
19. Mohseni S, Zarei N, Ragan ED. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *arXiv Preprint arXiv:1811.11839*; 2018.
20. Du M, Liu N, Hu X. Techniques for interpretable machine learning. *arXiv Preprint arXiv:1808.00033*; 2018.
21. Robinson C, Hohman F, Dilkina B. *A Deep Learning Approach for Population Estimation from Satellite Imagery*. New York: ACM; 2017: 47-54.
22. Zintgraf LM, Cohen TS, Adel T, Welling M. Visualizing deep neural network decisions: prediction difference analysis. *arXiv Preprint arXiv:1702.04595*; 2017.
23. Yang F, Liu N, Suhang W, Hu X. Towards interpretation of recommender systems with sorted explanation paths. *arXiv Preprint*; 2018.
24. Bussone A, Stumpf S, O'Sullivan D. *The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems*. New York: IEEE; 2015:160-169.
25. Kahng M, Andrews PY, Kalro A, Chau DHP. ActiVis: visual exploration of industry-scale deep neural network models. *IEEE Trans Vis Comput Graph*. 2018;24(1):88-97.
26. Goodall J, Ragan ED, Steed CA, et al. Situ: identifying and explaining suspicious behavior in networks. *IEEE Transactions on Visualization and Computer Graphics*. New York: IEEE Press; 2018.
27. Liu M, Shi J, Cao K, Zhu J, Liu S. Analyzing the training processes of deep generative models. *IEEE Trans Vis Comput Graph*. 2018; 24(1):77-87.
28. Kulesza T, Stumpf S, Burnett M, Yang S, Kwan I, Wong WK. *Too Much, Too Little, or Just Right? Ways Explanations Impact End Users' Mental Models*. New York: IEEE; 2013:3-10.
29. Glass A, McGuinness DL, Wolverton M. *Toward Establishing Trust in Adaptive Agents*: Conference on Intelligent User Interfaces, New York: ACM; 2008;13:227-236.
30. Bilgic M, Mooney RJ. *Explaining Recommendations: Satisfaction vs. Promotion*. Vol Beyond personalization workshop, IUI, 5; 2005:153. New York: ACM.
31. Berkovsky S, Taib R, Conway D. How to recommend?: user trust factors in movie recommender systems. *IUI '17*. New York, NY: ACM; 2017:287-300.
32. Kulesza T, Burnett M, Wong WK, Stumpf S. *Principles of Explanatory Debugging to Personalize Interactive Machine Learning*. New York: ACM; 2015:126-137.
33. Honeycutt D, Nourani M, Ragan E. Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. 2020;8:63-72.
34. Nourani M, Kabir S, Mohseni S, Ragan ED. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. 2019;7:97-105.
35. Krause J, Perer A, Ng K. *Interacting with Predictions: Visual Inspection of Black-Box Machine Learning Models*. New York: ACM; 2016: 5686-5697.
36. Nourani M, King J, Ragan E. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. 2020;8:112-121.

37. Nourani M, Roy C, Block JE, et al. *Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems*. International Conference on Intelligent User Interfaces, New York: ACM; 2021;26:340-350.
38. Allcott H, Gentzkow M. Social media and fake news in the 2016 election. *J Econ Perspect*. 2017;31(2):211-236.
39. Shu K, Sliva A, Wang S, Tang J, Liu H. Fake news detection on social media: a data mining perspective. *ACM SIGKDD Explorations Newslett*. 2017;19(1):22-36.
40. Rey C, Dugelay JL. A survey of watermarking algorithms for image authentication. *EURASIP J Appl Signal Process*. 2002;2002(1): 613-621.
41. Chakraborty A, Paranjape B, Kakarla S, Ganguly N. *Stop Clickbait: Detecting and Preventing Clickbaits in Online News Media*. New York: IEEE Press; 2016:9-16.
42. Sampson J, Morstatter F, Wu L, Liu H. *Leveraging the Implicit Structure Within Social Media for Emergent Rumor Detection*. New York: ACM; 2016:2377-2382.
43. Li Y, Gao J, Meng C, et al. A survey on truth discovery. *ACM Sigkdd Explorations Newslett*. 2016;17(2):1-16.
44. Shu K, Wang S, Liu H. Exploiting tri-relationship for fake news detection. *arXiv Preprint arXiv:1712.07709*; 2017.
45. Gupta A, Kumaraguru P, Castillo C, Meier P. *Tweetcred: Real-Time Credibility Assessment of Content on Twitter*. New York: Springer; 2014:228-243.
46. Seo H, Xiong A, Lee D. *Trust It or Not: Effects of Machine-Learning Warnings in Helping Individuals Mitigate Misinformation*. New York: ACM; 2019:265-274.
47. Gao M, Xiao Z, Karahalios K, Fu WT. To label or not to label: the effect of stance and credibility labels on readers' selection and perception of news articles. *Proc ACM Human Comput Interact*. 2018;2:1-16.
48. Gupta A, Lamba H, Kumaraguru P, Joshi A. *Faking Sandy: Characterizing and Identifying Fake Images on Twitter During Hurricane Sandy*. New York: ACM; 2013:729-736.
49. Vosoughi S, Roy D, Aral S. The spread of true and false news online. *Science*. 2018;359(6380):1146-1151.
50. Starbird K, Maddock J, Orand M, Achterman P, Mason RM. Rumors, false flags, and digital vigilantes: misinformation on twitter after the 2013 Boston marathon bombing. *IConference 2014 Proceedings*. Illinois: iSchools; 2014.
51. Starbird K, Arif A, Wilson T. Disinformation as collaborative work: surfacing the participatory nature of strategic information operations. *Proc ACM HumanComput Interact* 2019; 3(CSCW): 1–26.
52. Rashkin H, Choi E, Jang JY, Volkova S, Choi Y. Truth of varying shades: analyzing language in fake news and political fact-checking; 2017: 2931–2937.
53. Uscinski JE, Butler RW. The epistemology of fact checking. *Crit Rev*. 2013;25(2):162-180.
54. Mohseni S, Yang F, Pentyala S, et al. Machine learning explanations to prevent overtrust in fake news detection. *International AAAI Conference on Web and Social Media (ICWSM)*. Palo Alto, California: Association for the Advancement of Artificial Intelligence; 2021.
55. Vo N, Lee K. The rise of guardians: fact-checking url recommendation to combat fake news; 2018: 275–284.
56. Hansen C, Hansen C, Alstrup S, Grue Simonsen J, Lioma C. Neural check-worthiness ranking with weak supervision: finding sentences for fact-checking; 2019: 994–1000.
57. Wang X, Yu C, Baumgartner S, Korn F. Relevant document discovery for fact-checking articles; 2018: 525–533.
58. Yang F, Pentyala SK, Mohseni S, et al. *XFake: Explainable Fake News Detector with Visualizations*. New York: ACM; 2019:3600-3604.
59. Pennington J, Socher R, Manning C. *Glove: Global Vectors for Word Representation*; 2014:1532-1543.
60. Vaswani A, Shazeer N, Parmar N, et al. Advances in Neural Information Processing Systems. *Attention is All You Need*. New York: Curran Associates, Inc. 2017;30:5998-6008.
61. Chen T, Guestrin C. *Xgboost: A Scalable Tree Boosting System*. New York: ACM; 2016:785-794.
62. Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. *arXiv Preprint arXiv:1503.02531*; 2015.
63. Hullman J, Diakopoulos N. Visualization rhetoric: framing effects in narrative visualization. *IEEE Trans Vis Comput Graph*. 2011; 17(12):2231-2240.
64. Rubin VL, Conroy N. Discerning truth from deception: human judgments and automation efforts. *First Monday*. 2012;17(5):1-27.
65. Zuk T, Carpendale S. *Visualization of Uncertainty and Reasoning*. New York: Springer; 2007:164-177.
66. Greis M, Hullman J, Correll M, Kay M, Shaer O. *Designing for Uncertainty in HCI: When Does Uncertainty Help?* . New York: ACM; 2017:593-600.
67. Blair JP, Levine TR, Shaw AS. Content in context improves deception detection accuracy. *Human Communication Research*. 2010;36(3): 423-442.
68. Zubiaga A, Aker A, Bontcheva K, Liakata M, Procter R. Detection and resolution of rumours in social media: a survey. *ACM ComputSurv (CSUR)*. 2018;51(2):32.
69. Kurtz KJ, Miao CH, Gentner D. Learning by analogical bootstrapping. *J Learn Sci*. 2001;10(4):417-446.
70. Valdez AC, Ziefle M, Sedlmair M. Priming and anchoring effects in visualization. *IEEE Trans Vis Comput Graph*. 2017;24(1):584-594.
71. Wall E, Blaha LM, Paul CL, Cook K, Endert A. *Four Perspectives on Human Bias in Visual Analytics*. Springer: Springer; 2018:29-42.
72. Hullman J, Adar E, Shah P. *The Impact of Social Information on Visual Judgments*. New York: ACM; 2011:1461-1470.
73. Kittur A, Chi EH, Suh B. *Crowdsourcing User Studies with Mechanical Turk*. New York: ACM; 2008:453-456.
74. Suls R. Educational divide in vote preferences on track to be wider than in recent elections. *Pew Res Center*. 2016;9:15.
75. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access*. 2018;6:52138-52160.
76. Teevan J, Collins-Thompson K, White RW, Dumais ST, Kim Y. Slow search: information retrieval without time constraints; 2013: 1–10.

77. Norman D. *The Design of Everyday Things: Revised and Expanded Edition*. New York: Basic Books; 2013.
78. Leslie AM, Friedman O, German TP. Core mechanisms in "theory of mind". *Trends Cogn Sci*. 2004;8(12):528-533.

How to cite this article: Linder R, Mohseni S, Yang F, Pentyala SK, Ragan ED, Hu XB. How level of explanation detail affects human performance in interpretable intelligent systems: A study on explainable fact checking. *Applied AI Letters*. 2021;e49. doi:10.1002/ail2.49