Understanding the Role of Explanation Modality in AI-assisted Decision-making

VINCENT ROBBEMOND, Delft University of Technology, The Netherlands

OANA INEL*, University of Zurich, Switzerland

UJWAL GADIRAJU, Delft University of Technology, The Netherlands

Advances in artificial intelligence and machine learning have led to a steep rise in the adoption of AI to augment or support human decision-making across domains. There has been an increasing body of work addressing the benefits of model interpretability and explanations to help end-users or other stakeholders decipher the inner workings of the so-called "black box AI systems". Yet, little is currently understood about the role of modalities through which explanations can be communicated (*e.g.*, text, visualizations, or audio) to inform, augment, and shape human decision-making. In our work, we address this research gap through the lens of a credibility assessment system. Considering the deluge of information available through various channels, people constantly make decisions while considering the perceived credibility of the information they consume. However, with an increasing information overload, assessing the credibility of the information we encounter is a non-trivial task. To help users in this task, automated credibility assessment systems have been devised as decision support systems in various contexts (*e.g.*, assessing the credibility of news or social media posts). However, for these systems to be effective in supporting users, they need to be trusted and understood. Explanations have been shown to play an essential role in informing users' reliance on decision support systems. In this paper, we investigate the influence of explanation modalities on an AI-assisted credibility assessment task. We use a between-subjects experiment ($N = 375$), spanning six different explanation modalities, to evaluate the role of explanation modality on the accuracy of AI-assisted decision outcomes, the perceived system trust among users, and system usability. Our results indicate that explanations play a significant role in shaping users' reliance on the decision support system and, thereby, the accuracy of decisions made. We found that users performed with higher accuracy while assessing the credibility of statements in the presence of explanations. We also found that users had a significantly harder time agreeing on statement credibility without explanations. With explanations present, text and audio explanations were more effective than graphic explanations. Additionally, we found that the combinations of graphical and text and/or audio explanations were significantly effective. Such combinations of modalities led to a higher user performance than using graphical explanations alone.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; **User studies**; **Graphical user interfaces**; • **Information systems → Decision support systems**.

ACM Reference Format:

Vincent Robbemond, Oana Inel, and Ujwal Gadiraju. 2022. Understanding the Role of Explanation Modality in AI-assisted Decision-making. In *Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '22), July 4–7, 2022, Barcelona, Spain.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3503252.3531311

1 INTRODUCTION

Over the last decade, we have witnessed a surge in the adoption of AI-assisted decision-making across several domains [9, 18, 62], including critical domains like medical diagnoses [46], judicial sentencing [5], and hiring [66]. More recently, spurred by the interdisciplinary interest across research communities to help humans rely appropriately on AI

[*] Work partially performed while at Delft University of Technology, the Netherlands.

systems [23, 28], researchers in the field of explainable AI (XAI) have proposed different methods to use explanations and aid the interpretability of complex decision-support systems [39, 40, 42].

Concomitant with this growth in the adoption of AI systems is the constant need for people to make decisions based on the deluge of information we are exposed to [41]. The Web provides a plethora of information and continues to grow in size. Unfortunately, aside from a large amount of valuable information, the Web is also a source of false information. The rapid increase in fake news has become a widespread problem globally [37], and the diffusion of misinformation online has been shown to harm people's decision-making [68]. To make good decisions, we need to be able to assess or reflect on the credibility of the information we consume online.

Fact checking websites like Snopes[1] and Politifact[2] aim to provide reliable sources on the web, decreasing the fact-checking workload for individuals. Journalism deals with this at a professional level, producing and distributing information based on facts, albeit not devoid of biases [44]. Considering the ever-increasing stream of information produced on the internet, the task of finding and filtering information and assessing its credibility is challenging [33]. Scaling the credibility assessments is even harder, given the growth rate of misleading and false information being produced online [69]. To this end, tools have been developed to (partly) automate this process [1, 15, 24, 71]. Crowdsourcing has also helped address scalability issues regarding expert fact-checking. Recent works have demonstrated that adequately aggregated non-expert crowd worker assessments correlate well with expert assessments [35, 53].

To make automated credibility assessments comprehensible for end-users and stakeholders, prior work proposed using explanations alongside such assessments [50–52]. Yet, little is currently understood about the role of explanation modalities (*e.g.*, text, visualizations, or audio), to inform, augment, and shape human decision-making. Addressing this research gap, in this paper, we investigate the influence of explanation modality on AI-assisted decision-making through the lens of an automated credibility assessment system. This system assesses the credibility of statements (*i.e.*, or claims) made online and subsequently explains this assessment in different modalities. We explore different explanation modalities' effectiveness and users' trust in the automated credibility assessment system. We set out to answer the following research questions:

> **RQ1**: How do different explanation modalities corresponding to a credibility assessment system influence the perceived credibility of statements?
>
> **RQ2**: How do different explanation modalities corresponding to a credibility assessment system influence the perceived trust and user engagement with the system?

To answer our research questions, we designed and deployed a between-subjects user study ($N = 375$) with six conditions — a control condition in which users rate the credibility of a statement without AI-assisted support and five conditions in which users rate the credibility of a statement with AI-assisted support. More precisely, while performing the assessments in the latter cases, users were assisted by a credibility assessment system that included explanations in different modalities. Users were asked to label a statement on a scale from '*1: Not Credible*' to '*100: Credible*'. Additionally, we gathered information on the participants' affinity for interacting with technology, trust in, and usability of the credibility assessment system.

Our results indicate that explanations play a significant role in shaping users' reliance on the decision support system and, thereby, the accuracy of decisions made. We found that users, when being presented with explanations, had a higher degree of agreement and performed with a higher accuracy on the credibility assessment tasks. From the provided

[1] https://www.snopes.com/

[2] https://www.politifact.com/

explanation types, text and audio explanations significantly outperformed graphic explanations. Additionally, we found that combining graphic with text and/or audio explanations has a significant positive effect on user performance. To promote open science, we publicly share all our data and code.[3] Our findings have important implications for the broader XAI community and inform the design of explanations for future decision support systems.

2 BACKGROUND AND RELATED WORK

In this section, we review work on aspects related to credibility assessment and explanations. In Section 2.1, we discuss credibility assessment approaches. In Section 2.2, we discuss existing explanation-based approaches and their role in credibility assessment. Finally, in Section 2.3, we discuss information modalities to provide explanations to end-users.

2.1 Credibility Assessment

Credibility entails a multitude of aspects, among which, believability, trustworthiness, reliability, accuracy, fairness, and objectivity [3]. Thus, a credibility assessment can be described as an estimation of the trustworthiness or believability of something or someone. Research has shown, however, that people have difficulties in understanding and evaluating the veracity of the information they find online [20, 26, 70]. Thus, to support and augment human decision-making in terms of credibility assessment, typical solutions focus on making people reflect on the information they see by providing credibility markers [6, 30, 32, 49, 74].

In the news domain, Yaqub et al. [74] studied people's behavior in sharing news headlines with their social media peers. The news headlines were augmented with one in four credibility indicators (*i.e.*, when either fact-checkers, news media, public, or artificial intelligence techniques dispute the credibility of the headline). Their large-scale online experiment showed that fact-checking systems such as Snopes and Politifact are the most effective in decreasing sharing behavior of false information. A large body of research has also focused on assessing the credibility of tweets [6, 7, 13, 25]. For instance, Gupta et al. [25] introduced a semi-supervised ranking model using SVM-rank to label tweets with credibility scores in real-time. However, the system only uses the data available for the single tweet that is being assessed, disregarding historical or data relevant to the event mentioned in the tweet. Event level credibility assessment on Twitter usually looked into trending topics, which have gathered attention from many users within a relatively short time frame, *i.e.*, users create thousands of posts each minute [1].

Literature identified three main components that affect credibility *perception* of user generated content [3]: **context** (*i.e.*, environment, topic, and situation); available **features** of the content; and traits and cognitive heuristics of the **evaluator** (*i.e.*, topical knowledge and the selection of features in making a credibility judgement). According to AlMansour et al. [3], existing research focuses more on the content of a tweet and less on the context and evaluator. Thus, in this paper, we focus on these less explored factors to augment users with additional information regarding the tweet and help decide on its credibility. Furthermore, building on top of the results of Jahanbakhsh et al. [32], we also use fact-checking services such as Snopes and Politifact to select the tweets that we assess in our study.

2.2 Explanations

Automatically generated explanations supporting the results of machine learning models can help users better understand their output and be more receptive to them [29]. However, most users of machine learning models are by no means machine learning experts and can have trouble understanding the way these models work [56]. While some users

[3] https://osf.io/rdbz6/?view_only=97d7839f61774b8cb53f529694b45925

may be satisfied with consuming the results of models or receiving advice from AI systems, influencing their decision-making, others may want to know the rationale governing such advice to better understand the outcome — a sound understanding can make the model or the advice more likely to be accepted [17]. There is a growing demand for transparency in machine learning, as the application of these models is becoming increasingly common, while at the same time, models are becoming more complex and are gaining influence [43]. Explanations can provide this transparency through model-agnostic frameworks that increase understanding of black-box models [36], providing input evidence by highlighting words that are key to the decision [38], or generating automated (natural language) rationale in real-time [16].

Nunes and Jannach [45] performed a systematic review of explanations in decision support and recommender systems and derived a taxonomy of explanations. Based on previously proposed explanation purposes in [60], three explanation objectives were derived regarding stakeholder goals (*i.e.*, acceptance intention, use intention, among others), user-perceived quality factors (*i.e.*, confidence, ease of use, perceived transparency, among others), and explanation purposes (*i.e.*, effectiveness, efficiency, transparency, among others). This work focuses on four aspects: trust, satisfaction, efficiency, and effectiveness. Trust and satisfaction are user-perceived quality factors — trust refers to the system being perceived as trustworthy, and satisfaction refers to usefulness and usability [60]. Efficiency and effectiveness are explanation purposes, with efficiency pertaining to helping users make decisions faster and effectiveness to assisting users in making good decisions.

Furthermore, the work by Nunes and Jannach [45] and Tintarev and Masthoff [61] showed that the result, its decisive features (most influential features for the given result), and the confidence (in the result) are associated with better performance of explanations. To this end, the explanations provided in our task consist of: (1) indicating whether the statement is believed to be credible or not credible (*i.e.*, result); (2) showing the aggregated sentiment and stance of Web articles on this statement, and the article attention words (*i.e.*, decisive features), and (3) showing the percentage of true/false and Web source credibility (*i.e.*, confidence).

A few credibility assessment tools employ explanations as part of their user experience. The FeedReflect tool [7] is used as a browser extension that provides visual cues (content highlighting and dimming content from non-mainstream sources) to nudge people to reflect on the feed items credibility. Popat et al. [50] automated the assessment of credibility in emerging claims on the internet while providing suitable, user-interpretable explanations from selected sources. The CredEye system [51] assesses the credibility of a claim by analyzing relevant Web articles. Explanations are provided as snippets from considered Web sources combined with their trustworthiness. The neural network model DeClarE [52] aggregates signals from external evidence articles to assess the credibility of natural language claims and generates explanations by means of decisive features. However, to the best of our knowledge, these systems and explanations have not been assessed from the users' perspective. Thus, in this paper, we study the influence of the explanations generated by the CredEye [51] and DeClarE [52] tools on the accuracy of human decision-making for credibility assessment.

2.3 Information and Explanation Modality

Information modality refers to how information is presented and, consequently, its effect on how people process this information. In the context of decision-support systems such as recommender systems or classification models, the most common explanation modalities are textual and visual (graphical) [8, 31, 54, 59, 63, 64, 72, 73]. Tran et al. [63] used textual explanations to justify recommended restaurants to groups of people. Yang et al. [73] studied user's perception of trust in a classification model when augmented with various visualization designs and found that while each visual explanation increased user trust, they could also persuade users to accept wrong classification outcomes.

Tsai and Brusilovsky [64] compared three textual explanations and twelve visual explanations for three similarity-based people recommendation models. They found that visual explanations are preferred over textual ones and lead to better representation of explanation goals, such as scrutability and persuasiveness, among others.

A long-standing line of research in this area is the Cognitive Load Theory (CLT) and the Modality Effect [57, 58]. According to the available models of multimedia learning, cognitive processing of related text and pictures involves selecting and organizing the relevant elements of visual and auditory information. The result is a coherent, unified representation of all aspects processed in the learner's working memory. In essence, CLT argues that limited working memory can be effectively expanded by using more than one presentation modality [57, 58].

Cao et al. [11, 12] studied the effect of different modalities (*i.e.*, text, image, speech, and sound) on people's cognitive load and performance in a high-load information presentation scenario. The users played the role of crisis managers after an earthquake. Their task was to communicate the location of victims to rescue workers. The experiment showed that combining text and speech modalities provides the optimal way to present information. In a study on multimodal and interactive explanations in the context of visual question answering (VQA) [2], the authors showed that participants' prediction accuracy improved significantly in the presence of explanations when the system was incorrect. Furthermore, participants' explanations ratings indicated their effectiveness in an AI-assisted human-machine collaboration task. Similarly, Park et al. [48] found that visual and textual explanations generated by a VQA model were complementary, and in some cases, visual indicators were more explanatory than textual ones and vice versa. Additionally, the authors showed that providing explanations enables humans to assess more accurately whether a system assessment is correct.

The advantages of combining modalities in information presentation have also been shown in the context of in-vehicle information systems [10]. Participants scored better at driving and secondary tasks, had faster reaction times, and lower cognitive load while presented with multimodal information. The study affirmed the advantages of combining modalities, such as enhanced communication robustness due to redundant or complementary use of modalities. Similarly, Szymanski et al. [59] showed that even though study participants preferred graphical explanations, their performance in correctly identifying the reading time of a news article was better when augmented with textual explanations.

To the best of our knowledge, audio explanations have not been explored in the context of decision support systems. However, work in both Cognitive Load Theory and multimedia principles for learning showed that graphics combined with audio designs performed better than graphics combined with text. Thus, based on this insight, we experiment with several explanation modalities (text, graphics, audio) and combinations of those (graphics + text, graphics + audio).

3 STUDY DESIGN

The goal of our study is to understand the effects different explanation modalities have on users' perceived credibility of online statements. To achieve this goal and address our research questions, we conducted a between-subjects user study. Thus, in this section, we describe the data (*i.e.*, the statements or claims), the procedure, and the measures of the study.

3.1 Task data

For our study, we curated 40 statements (*i.e.*, claims) from the training sets of the CredEye [51] and DeClarE [52] credibility assessment systems, which have been gathered from Snopes and PolitiFact. Each statement was labeled with a credibility label or stance. We selected the statements such that they were equally divided into four credibility bins, each credibility bin corresponding to a credibility label, *i.e.*, *not credible*, *somewhat not credible*, *somewhat credible*, and *credible*, resulting in ten statements per credibility bin. We cluster the selected statements into these four bins because it has been shown that more coarse-grained truthfulness scales are preferred in crowdsourcing settings [35]. The list of all selected statements can be checked in our repository[4].

Each statement was assigned a set of values, in line with the credibility bin it belongs to (ground-truth). These values are the parameters used for building the explanations. The following parameters are used in the explanations, based on existing research (see Section 2.2) and the output of the CredEye [51] and DeClarE [52] credibility assessment systems:

- *credibility bin or credibility label*: not credible/somewhat not credible/somewhat credible/credible - systems classification regarding the statement;
- *credibility percentage*: [0:25, 26:50, 51:75, 76:100]% - refers to the system probability of the statement to be not credible, somewhat not credible, somewhat credible, or credible;
- *number of articles considered, number of supporting articles, and number of opposing articles*: refers to the total number of web articles that are consulted to check the credibility of the statement, and how many out of these support or oppose the statement;
- *average source credibility*: [1:100] - indicates the average credibility rating of the consulted articles.

Thus, we design our explanations based on the template below (explanation parameters are shown in italics between brackets). An example of such an explanation is shown in Figure 1.

> The system believes this claim to be <*credibility bin/label*>.
>
> According to consulted web-sources the probability of this claim to be true is <*credibility percentage*>. In total <*number of articles considered*> articles were considered of which <*number of supporting articles*> indicating this statement is credible and <*number of opposing articles*> indicating this statement is not credible. The consulted sources have an average credibility rating of <*average source credibility*>%.

3.2 Independent variables

We have a single independent variable, the explanation modality, with six conditions: text, audio, graphical, combination 1 (text + graphical), combination 2 (audio + graphical), and no explanation (*i.e.*, a control group for which no explanation is provided).

The *text explanation* is based on the template shown in Section 3.1. Figure 1 shows a screenshot of the task user interface that users in this condition received.

The *audio explanation* is generated by first generating the text explanation for that statement the exact same way as for the text explanation described above, and then using the Mozilla TTS [5] tool to generate the audio version of the text explanation. Mozilla TTS is an open-source Text-to-Speech tool that provides pre-trained, high-quality models. We selected the pre-trained Tacotron2 model with the LJSpeech dataset (English), which has good performance on both long and short sentences, given that our explanations contain a mix of those. Figure 3 shows a screenshot of the task user interface that users in this condition received.

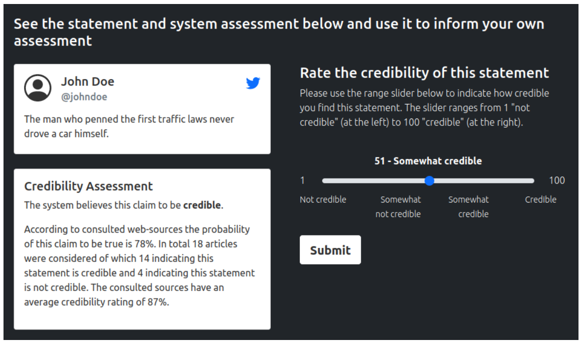
The *graphical explanation* is generated by using the same credibility assessment parameters (shown in Section 3.1) of the statement to generate graphs with the ChartJS library.[6] The graphs were designed with accessibility in mind, using

[4] https://osf.io/rdbz6/?view_only=97d7839f61774b8cb53f529694b45925
[5] https://github.com/mozilla/TTS
[6] https://github.com/chartjs/Chart.js

a colorblind-friendly color scheme. The explanation consists of three components; a bar chart depicting the *credibility percentage*; a pie chart depicting the *number of articles considered, number of supporting articles, and number of opposing articles*; and a bar chart depicting the *average source credibility*. Figure 2 shows the graphical explanation which was shown to the participants in this condition.

The *text+graphical condition* consists of the exact contents of the text and graphic explanations and was created by combining the text and graphic components described above.

The *graphical+audio condition* consists of the exact contents of the audio and graphic explanations and was created by combining the audio and graphic components described above.



See the statement and system assessment below and use it to inform your own assessment

John Doe
@johndoe

The man who penned the first traffic laws never drove a car himself.

Credibility Assessment

The system believes this claim to be **credible**.

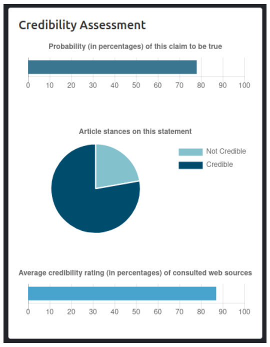
According to consulted web-sources the probability of this claim to be true is 78%. In total 18 articles were considered of which 14 indicating this statement is credible and 4 indicating this statement is not credible. The consulted sources have an average credibility rating of 87%.

Rate the credibility of this statement

Please use the range slider below to indicate how credible you find this statement. The slider ranges from 1 "not credible" (at the left) to 100 "credible" (at the right).

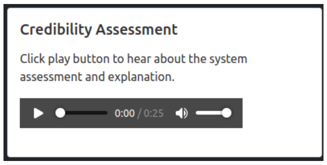
51 - Somewhat credible

Fig. 1. UI text explanation.



Credibility Assessment
Probability (in percentages) of this claim to be true
0
10
20
30
40
50
60
70
80
90
100
Article stances on this statement
Not Credible
Credible
Average credibility rating (in percentages) of consulted web sources
0
10
20
30
40
50
60
70
80
90
100

Fig. 2. UI graphic explanation.



Credibility Assessment

Click play button to hear about the system assessment and explanation.

Fig. 3. UI audio explanation.

3.3 Measured Variables

Statement credibility. The participants rate the credibility of each statement on a range from *1: Not Credible* to *100: Credible*. We used this scale to align with the output of the credibility assessment tools.

Affinity for technology. The Affinity for Technology Interaction (ATI) scale was used to assess user propensity towards interacting or engaging with technology. The 9-item ATI questionnaire is seen as "a core personal resource for users' successful coping with technology" [4, 21]. Each item was annotated on a Likert scale, from *1: Completely Disagree* to *6: Completely Agree.*

User engagement. We used the User Engagement Scale (UES) questionnaire to measure self-reported user engagement with the credibility assessment task interface across various dimensions [47]. We used the short version of the questionnaire, UES Short Form, to evaluate the following factors: (1) the focused attention (feeling absorbed in the interaction and losing track of time) and (2) the perceived usability (negative effect experienced as a result of the interaction and the degree of control and effort expended). We evaluated these two factors because they are directly related to the explanation goals we are interested in, namely satisfaction, efficiency, and effectiveness, for which we are evaluating the different explanation modalities. The questionnaire consists of a 6-item list annotated on a Likert scale from *1: Strongly Disagree* to *5: Strongly Agree*.

User trust in automation. We used the Trust in Automation (TiA) questionnaire [34] to measure the level of trust in an automated solution (in this case, the system's credibility assessment). It consists of 19 items annotated on a Likert scale from *1: Strongly Disagree* to *5: Strongly Agree*.

3.4 Task Setup and Procedure

We used the Prolific[7] crowdsourcing platform to publish our task and collect data. When participant decided to take part in our study by clicking the "Open study link in a new window" button, they would be taken to a website where we deployed our task as a web application. The link contains a unique and anonymous identifier for the participant and session, to identify successful submissions.

The participants were first greeted with a brief introductory text explaining their task and an informed consent. Then, participants were first asked to answer three non-mandatory demographic questions related to their *age*, *gender*, and *education level* and rate the items in the *Affinity for Technology Interaction* (ATI) scale.

Upon completing the ATI questionnaire, the participants were asked to assess the *credibility* of four statements. For each participant, the statements were randomly selected from our dataset (see Section 3.1, one statement from each credibility bin. Then, each participant was assigned to one of our six conditions (control group, text explanation, audio explanation, graphic explanation, text+graphic explanation or audio+graphic explanation). For each statement, the participant had to rate the credibility of the statement by setting a slider (with steps of 1) to a value between 1 (not credible) and 100 (credible). A single statement was presented at a time, with the next one showing after the previous assessment was submitted. The statements appeared in random order for each participant.

After rating the credibility of the statements, the participants were asked to complete the *User Engagement Scale* (UES) and then the *Trust in Automation* (TiA) questionnaire. Finally, the participants were shown the completion page and given the option to submit their submissions. When the submit button was clicked, participants were taken back to the Prolific page with a completion code marking the successful completion of the task.

To ensure submission quality, three attention checks were included in the task, one each in the ATI, UES, and TiA scales/questionnaires. These attention checks were questions indicating they were attention checks and telling the participant which answer to provide.

3.5 Participants

We calculated the required sample size (*N=324*) by performing an apriori power analysis using the G*Power [19] tool. To ensure high-quality study submissions, we required participants on the Prolific platform with an approval rate of at least 90% and English as primary language. In total, 418 people participated in our study, and 375 were approved (24 stopped before completing and 19 were rejected). Approved participants received payment in line with £7.56/hour.

[7] https://www.prolific.co/

3.5.1 *Demographics of approved participants.*

Of the participants, 292 identified as female, 71 as male, and 12 as other. Participants are distributed across all age ranges, but the majority of the participants are between 18 and 27 years old, namely 233 (62%) (28-37: 86, 38-47: 36, 48-57: 13, 58+: 5, no answer: 2). The majority of our participants had either a high school (37%) or a bachelor's degree (39%). A smaller fraction of participants had a master's degree (13%) or vocational training (6%), while 5% had either less than a high school degree, a doctoral degree, or chose not to respond.

3.6 Pilot

A pilot study was set up with the following goals in mind: (1) to determine the time needed for participants to complete the survey and set up appropriate payment [14], (2) to receive early participant feedback [22], and (3) to test our task system. The pilot experiment procedure followed the procedure in Section 3.4. A total of 10 participants (2 per condition) with English as first language were recruited on Prolific. All 10 participants successfully completed the study without any errors, so no significant changes to the task or system were deemed necessary. To make the system clearer to interact with, small changes regarding the presentation of the user controls on the task interface were made.

3.7 Statistical Tests and Analysis

During initial data exploration, we perform linear regressions on metrics of interest. We used the non-parametric Kruskal-Wallis test because our data fails the normality assumption. A significant test result for the Kruskal-Wallis test only indicates whether at least one sample is statistically significantly different than the rest, but does not indicate which. To identify the statistically significant differences, we perform a post-hoc analysis using the Mann-Whitney U test. We apply the Bonferroni and Holm-Bonferroni corrections (with $p < \alpha/m$ being $p < 0.0033$, where m refers to the number of repeated measures, namely 15) to account for repeated measures. For our statistical analysis we used the NumPy [27], SciPy [67], statsmodels [55], and Pingouin [65] Python packages.

4 RESULTS

We now present the results of our study. Participants were assigned an explanation modality at random, in a balanced manner (we had between 57 and 80 participants per condition, with the maximum number of participants in the control condition) (see Figure 4 for exact numbers). We recall here that perceived credibility was measured on a scale from 1 to 100, divided into 4 bins: "not credible" [1,25], "somewhat not credible" [26,50], "somewhat credible" [51,75], and "credible" [76,100].

The statements used in our study were balanced with respect to their credibility bins. One would, therefore, expect the mean credibility score across all tasks to be around 50. However, as shown in Figure 4a, we found that the highest mean credibility score across all explanation modalities was well below 50. This suggests an overall tendency for participants in our study to label the statements as less credible across all conditions. We also observed that the audio explanation modality corresponded to the highest average credibility rating, 43.9, while the graphic modality corresponded to the lowest average credibility modality.

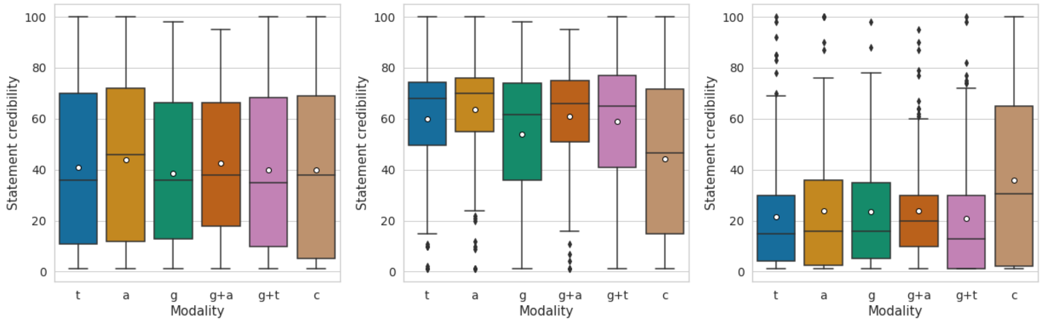
Looking at the accuracy of decisions made by users on assessing the statement credibility, the percentages of correctly labeled statements for each modality were found to be: text 47.2%; audio 55.6%; graphic 38.6%; graphic+text 49.6%; graphic+audio 55.3%; control 34.1%. We carried out multiple chi-square tests of independence to examine the relation between the explanation modalities and the accuracy of the decision reached. The results from the corresponding chi-square tests are presented in Table 1.

Modality	$\tilde{\chi}^2$ **Statistic**	$p - value$
t vs. c	9.49	0.002
a vs. c	24.57	$\mathbf{7.15}e^{-07}$
g vs. c	0.99	0.32
g+a vs. c	23.57	$\mathbf{1.20}e^{-06}$
g+t vs. c	13.19	**0.0002**

Table 1. Results from chi-square tests of independence to examine the relation between different explanation modalities with respect to the control condition. Statistically significant p-values after Bonferonni correction are indicated in **bold** for $p < .001$.

Credibility	**Effect**	**Significance**
Credible	text > control	p < 0.05
	audio > control	p < 0.005
	g+a > control	p < 0.01
	g+t > control	p < 0.05
Somewhat credible	audio > control	p < 0.001

Table 2. Significant effects of explanation modality on perceived credibility.



(a) All statements

(b) (Somewhat) Credible

(c) (Somewhat) Not credible

Fig. 4. Boxplots of credibility values given by participants, grouped by explanation modality. Means shown as white dots and medians shown as horizontal lines inside the boxes. t: text ($N = 62$), a: audio ($N = 58$), g: graphic ($N = 57$), g+a: graphic+audio ($N = 57$), g+t: graphic+text ($N = 61$), c: control ($N = 80$).

We found a significant difference in the assessments between the explanation modalities used for the "Credible" ($p < 0.0005$), "Somewhat credible" ($p < 0.01$), and "Not credible" ($p < 0.05$) statement groups using Kruskal-Wallis tests (see Table 2).

Figure 4c depicts the credibility ratings for the statements in the overall not credible range [1-50] and Figure 4b for those in the overall credible range [51-100]. Aside from the control group, we observe a clear division in credibility scores between credible and not credible statements. The box spread for the control group in both Figures 4c and 4b indicate users had a harder time agreeing on the credibility of statements when explanations were not provided.

The results of the Kruskal-Wallis test for the overall credible statements group (separated by condition) (H-statistic 38, p < 3.7e-7) and the overall not credible group (separated by condition) (H-statistic 17.5, p < 0.005) indicated significant differences exist between the explanation types. Post-hoc analysis with the Mann-Whitney U test revealed significant differences with the control group for: audio ($p < 0.000001$, Cohen's d 0.68), text ($p < 0.0005$, Cohen's d 0.55), graphic+audio ($p < 0.0005$, Cohen's d 0.6), and graphic+text ($p < 0.005$, Cohen's d 0.51). Only the graphic explanations did not result in significantly better credibility assessments. In addition, the audio explanations significantly outperformed the graphic explanations ($p < 0.05$, Cohen's d 0.39).

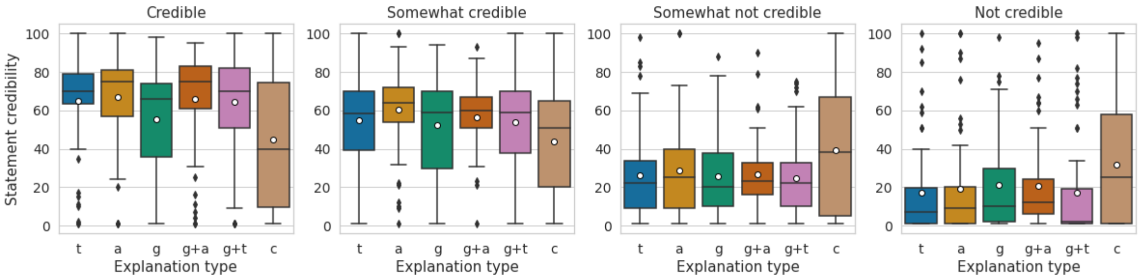


Fig. 5. Boxplots of credibility values given by participants for the different credibility bins, grouped by explanation modality. Means shown as white dots and medians shown as horizontal lines inside the boxes.

For the overall not credible statements, only the graphic+text explanations significantly improved credibility assessments compared to the control group ($p < 0.05$, Cohen's d 0.51), a notable result is text explanations ($p = 0.55$, Cohen's d 0.49).

4.1 Agreement with the System

Initial analysis shows there is a positive correlation between system credibility assessment and user-perceived credibility in the presence of explanations. The control group showed the same correlation, although to a much lesser extent. User perceived credibility also converged with system assessment as users were exposed to more statements/explanations, *i.e.*, users tended to have a higher chance of agreeing with the system the more statements they were rating. The overall accuracy of the users' decision while rating the first statement was 35.5%, rising to 63.2% by the fourth statement. This phenomenon was strongest for users who initially were more skeptical of the system, *i.e.*, their first assessment differed a lot from the system, while by their fourth assessment, they tended to mostly agree.

4.2 Perceived Credibility

We now analyze whether the explanation modality influences perceived credibility. We compare user assessments using the Kruskal-Wallis test, Bonferroni corrected for multiple comparisons. We look at each credibility bin separately, starting with the "credible" bin. The H-statistic of 24.7 with a p-value of 0.00015 indicates there is a significant difference between the explanation types. An overview of the data can be found in the first boxplot of Figure 5.

Both the mean and median statement credibility are the lowest in the control group, falling in the "somewhat not credible" bin, indicating that without an explanation present, credible statements are more likely to be perceived as not being credible. In contrast, credible statements with audio explanations present have the highest credibility rating at 66.7, with text, graphic+text, graphic+audio very close. For all explanation types, the lowest credibility rating given is 1, and as stated earlier, both graphic and graphic+audio have not scored above 98 and 95, respectively. Additionally, the average credibility rating for none of the explanation types is in the "credible" bin, they are all below 76, with the average being the lowest for the group without an explanation.

When performing post-hoc analysis with the Mann-Whitney U test, these findings were solidified with a significant difference shown between the control group and the text ($p < 0.002$, Cohen's d 0.66), audio ($p < 0.005$, Cohen's d 0.71), graphic+text ($p < 0.05$, Cohen's d 0.62), and graphic+audio ($p < 0.01$, Cohen's d 0.67) groups.

Moving on to the somewhat credible bin, for which the data overview can be found in the second boxplot displayed in Figure 5. All group means are within the ground-truth credibility bin (somewhat credible), except for the control group, which has a mean credibility score in the somewhat not credible group. This indicates that explanations are helping users make more accurate decisions for somewhat credible statements, which is in line with the accuracy numbers mentioned at the beginning of this section. The Kruskal-Wallis results for the somewhat credible statements grouped by explanation type are H-statistic of 15.7 with a p-value of 0.008, indicating there is a significant difference between the explanation types.

Comparing the credible and somewhat credible graphs in Figure 5, we note a few differences. The credibility scores for the text explanations have a lower mean and median, which is to be expected as the ground-truth expects scores on the interval [51-75] while the credible bin interval is at [76-100]. However, for the somewhat credible statements, the spread in credibility scores also seems larger, indicating users were less sure about the credibility of somewhat credible statements vs. the credible statements. The inverse, however, seems true for the audio explanations and the control group, where there is a smaller spread on the somewhat credible statements. The graphic, g+a, and g+t explanations show similar data on both the credible and somewhat credible, albeit with a slightly lower mean/median for the somewhat credible statements (which is to be expected).

Post-hoc analysis with the Mann-Whitney U test revealed that only the audio explanations were significantly different from the control group ($p < 0.005$, Cohen's d 0.66).

The Kruskal-Wallis results for the somewhat not credible and not credible statements grouped by explanation type indicated there is no statistical significance difference between groups (somewhat not credible: H-statistic = 8.8, p = 0.12, not credible: H-statistic = 13.4, p = 0.02).

4.3 Trust in Automation

Initial analysis shows there is a positive correlation between user agreement with the system and their trust scores. Users with a higher trust score in the TIA questionnaire converged with the system assessments. For the control group, however, the correlation is less stronger.

Next, we applied the Kruskal-Wallis test on the questionnaire answers separated in 6 groups (5 modality types and 1 control group). The results indicate that there is no statistically significant difference between the explanation types and the total TiA score (H-statistic = 7.05, p-value = 0.22). We then did an analysis of only the trust score component of the TIA questionnaire. Again, we found no statistically significant difference between the modality groups (H-statistic = 6.56, p-value = 0.26). These results indicate that users' trust is not affected by the explanation modality.

4.4 User Satisfaction, Efficiency, and Effectiveness

We analyzed the results of the User Engagement questionnaire for user satisfaction, efficiency, and effectiveness. Our analysis shows that there is a positive correlation between user agreement with the system and their UE scores, but only for audio explanations. The control group, along with the text, graphic, and graphic+text groups showed a slight negative correlation in this regard, with no apparent correlation for the graphic+audio explanation.

We found no statistically significant difference between the explanation types and the total User Engagement score (*H-statistic* = 3.93, *p*-value = 0.56). Performing the Kruskal-Wallis test on the perceived usability component resulted in an *H-statistic* of 2.5 with a *p*-value of 0.78, again showing no significant difference between the modality groups.

5 DISCUSSION

We studied the role of explanation modalities in informing, augmenting, and shaping human decision-making in a credibility assessment setting. The overall results show that user accuracy increases significantly in a credibility assessment setting when explanations are provided.

5.1 Perceived Credibility

Our main focus was to study the influence of explanation modalities on the accuracy of user decisions on credibility. Addressing our first research question of *how different explanation modalities corresponding to a credibility assessment system influence the perceived credibility of statements*, we first separated the statements in two equally divided credibility bins, namely the credible bin (range [51,100]) - consisting of the *credible* and *somewhat credible* statements, and the not credible bin (range [1,50]) - consisting of the *not credible* and *somewhat not credible* statements. The graphs in Figure 4 shows a clear distinction between the *credible* and *not credible* statements for the users assisted by explanations. The control group, however, looks very similar, with only a slight shift up or down with respect to the ground truth credibility, which indicates that the presence of explanations has a positive effect on the overall users' decision accuracy.

The text, audio, graphic+audio, and graphic+text modalities did not have people agree with the credible ground truth on average. However, they significantly outperformed both graphic and the control group. The control group showed participants were mostly leaning toward labeling the credible statements as being somewhat not credible, again demonstrating the impact of explanations.

Another thing to note is the fact that the graphic explanations were significantly outperformed by the other explanations (except for "not credible" statements), showing the importance of adding another modality to graphic explanations. However, the combinations of graphic+text and graphic+audio had no significant difference to the modality added to the graphic explanation (text and audio). A possible explanation for graphic+audio outperforming the graphic modality could be an increased trust in the system.

For the statements in the credible bin, the median answer value in the control group falls into the "somewhat not credible" bin, while the medians of all treatments that included an explanation fell into the "somewhat credible" bin. The control groups for the other credibility bins were already leaning toward their respective ground-truth credibility bin. This credible bin is also the bin where we almost exclusively see significant results when zooming in with pairwise post-hoc tests, possibly explained because here there was more room for improving the accuracy.

5.2 Perceived Trust and Engagement

We also explored the influence of explanation modalities on the users' perceived trust in the system and their engagement with the system, addressing our second research question. Our results showed that the presence of audio explanations increased users' perceived trust and engagement with the system. Consequently, this led to an increased agreement with the system's assessment. Further data analysis, however, revealed that there is no statistically significant difference for user trust and engagement in our decision support system, based on explanation modality. This could be explained by the results found by Wang and Yin [72], who found that the explanation that is considered to resemble how humans explain decisions (*i.e.*, counterfactual explanation) does not seem to improve calibrated trust. A way to test this is by measuring three levels of support for participants assessing statement credibility: without a decision-support system; with a decision-support system but without explanations; with a decision-support system with explanations. Research done by Yang et al. [73] suggests that users without a decision-support system perform worse than users being supported by a decision-support system.

5.3 Limitations and Future Work

We have identified several limitations of our approach. First of all, in our experimental setup, we did not study the influence of explanation modalities in settings where the AI prediction is incorrect. Second, we did not measure the user perceived credibility in setups where the AI prediction is given without being augmented with explanations. However, since our goal was to investigate whether different explanation modalities improve user decision accuracy compared to the control group (where no explanation is given), we only used correct predictions and augmented them with explanations. Future studies could focus on studying the influence of explanations in settings where the predictive model is wrong and mislabels statements' credibility. Such studies would also allow us to investigate how people's perceived trust changes when the credibility assessment tool has unpredictable accuracy. In such cases, more fine-grained assessments of user trust and user engagement (*i.e.*, after rating the credibility of each statement) would be needed. Third, we studied the influence of explanation modalities on a single, straightforward task, namely credibility assessment. Furthermore, our study participants were fairly educated, which could influence their assessments. We argue that future studies could study the influence of various demographic characteristics on human-perceived credibility.

6 CONCLUSION

In this paper, we investigated the influence of explanation modality on AI-assisted decision-making, leveraging a credibility assessment system as a lens for our exploration. This allowed us to better understand end-users' perceived credibility and decision accuracy when relying on the credibility assessment system. We designed and motivated our study following principles regarding information modality from the Cognitive Load Theory. Our results indicate that explanations have a significantly positive effect on user performance in assessing the credibility of statements. This is consistent with prior works that have explored the potential of explanations. However, our work expands into understanding the role of modalities of explanations in shaping human decision-making. We found that text and audio explanations were the most effective in increasing users' accuracy in assessing statements' credibility. Additionally, graphical explanations were only effective when combined with either text or audio explanations. In relation to decision accuracy, we found that the accuracy of user decisions increased as they made more assessments (measured across the 4 different decisions each user made).

Users reported increased trust levels when explanations were present and tended to have a higher agreement with the accurate system assessments, especially for the combination of graphic and audio explanations. Audio explanations were the only mode of explanation for which a positive correlation was found between user perceived engagement/usability and system agreement. Our work has important implications for the broader explainable AI (XAI) community, and our findings can inform the future design of AI systems that aim to augment human decision-making.

REFERENCES

[1] M. Al-Qurishi, M. S. Hossain, M. Alrubaian, S. M. M. Rahman, and A. Alamri. 2018. Leveraging Analysis of User Behavior to Identify Malicious Activities in Large-Scale Social Networks. *IEEE Transactions on Industrial Informatics* 14, 2 (2018), 799–813.

[2] Kamran Alipour, Jürgen P. Schulze, Yi Yao, Avi Ziskind, and Giedrius Burachas. 2020. A Study on Multimodal and Interactive Explanations for Visual Question Answering. *CoRR* abs/2003.00431 (2020). arXiv:2003.00431 https://arxiv.org/abs/2003.00431

[3] Amal Abdullah AlMansour, Ljiljana Brankovic, and Costas S. Iliopoulos. 2014. Evaluation of Credibility Assessment for Microblogging: Models and Future Directions. In *Proceedings of the 14th International Conference on Knowledge Technologies and Data-Driven Business* (Graz, Austria) *(i-KNOW*

'14). Association for Computing Machinery, New York, NY, USA, Article 32, 4 pages. https://doi.org/10.1145/2637748.2638439

[4] Christiane Attig, Daniel Wessel, and Thomas Franke. 2017. Assessing Personality Differences in Human-Technology Interaction: An Overview of Key Self-report Scales to Predict Successful Interaction. In *HCI International 2017 – Posters' Extended Abstracts*, Constantine Stephanidis (Ed.). Springer International Publishing, Cham, 19–29.

[5] Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. 2018. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Conference on fairness, accountability and transparency*. PMLR, 62–76.

[6] Md Momen Bhuiyan, Michael Horning, Sang Won Lee, and Tanushree Mitra. 2021. NudgeCred: Supporting News Credibility Assessment on Social Media Through Nudges. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–30.

[7] Md Momen Bhuiyan, Kexin Zhang, Kelsey Vick, Michael A Horning, and Tanushree Mitra. 2018. FeedReflect: A tool for nudging users to assess news credibility on twitter. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 205–208.

[8] Carrie J Cai, Jonas Jongejan, and Jess Holbrook. 2019. The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces*. 258–262.

[9] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[10] Yujia Cao. 2010. The Use of Modality in In-Vehicle Information Presentation: A Brief Overview. In *Proceedings of the 2nd International Workshop on Multimodal Interfaces for Automotive Applications* (Hong Kong, China) *(MIAA '10)*. Association for Computing Machinery, New York, NY, USA, 6. https://doi.org/10.1145/2002368.2002371

[11] Yujia Cao, Mariët Theune, and Anton Nijholt. 2009. Modality Effects on Cognitive Load and Performance in High-Load Information Presentation. In *Proceedings of the 14th International Conference on Intelligent User Interfaces* (Sanibel Island, Florida, USA) *(IUI '09)*. Association for Computing Machinery, New York, NY, USA, 335–344. https://doi.org/10.1145/1502650.1502697

[12] Yujia Cao, Mariët Theune, and Anton Nijholt. 2010. *Cognitive-Aware Modality Allocation in Intelligent Multimodal Information Presentation.* Springer London, London, 61–83. https://doi.org/10.1007/978-1-84996-507-1_3

[13] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. 675–684.

[14] Justin Cheng, Jaime Teevan, and Michael S Bernstein. 2015. Measuring crowdsourcing effort with error-time curves. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1365–1374.

[15] Nadia Conroy, Victoria Rubin, and Yimin Chen. 2015. Automatic Deception Detection: Methods for Finding Fake News.

[16] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. 2019. Automated Rationale Generation: A Technique for Explainable AI and Its Effects on Human Perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 263–274. https://doi.org/10.1145/3301275.3302316

[17] Alexander Erlei, Richeek Das, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2022. For What It's Worth: Humans Overwrite Their Economic Self-interest to Avoid Bargaining With AI Systems. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '22)*. https://doi.org/10.1145/3491102.3517734

[18] Alexander Erlei, Franck Nekdem, Lukas Meub, Avishek Anand, and Ujwal Gadiraju. 2020. Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 8. 43–52.

[19] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39 (2007), 175–191.

[20] Julian Fraillon, John Ainley, Wolfram Schulz, Tim Friedman, and Daniel Duckworth. 2020. *Preparing for life in a digital world: IEA International computer and information literacy study 2018 international report.* Springer Nature.

[21] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. *International Journal of Human–Computer Interaction* 35, 6 (2019), 456–467. https://doi.org/10.1080/10447318.2018.1456150 arXiv:https://doi.org/10.1080/10447318.2018.1456150

[22] Ujwal Gadiraju, Jie Yang, and Alessandro Bozzon. 2017. Clarity is a worthwhile quality: On the role of task clarity in microtask crowdsourcing. In *Proceedings of the 28th ACM conference on hypertext and social media*. 5–14.

[23] Akshit Gupta, Debadeep Basu, Ramya Ghantasala, Sihang Qiu, and Ujwal Gadiraju. 2022. To Trust or Not To Trust: How a Conversational Interface Affects Trust in a Decision Support System. In *Proceedings of the ACM Web Conference*.

[24] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. TweetCred: A Real-time Web-based System for Assessing Credibility of Content on Twitter. *CoRR* abs/1405.5490 (2014). arXiv:1405.5490 http://arxiv.org/abs/1405.5490

[25] Aditi Gupta, Ponnurangam Kumaraguru, Carlos Castillo, and Patrick Meier. 2014. Tweetcred: Real-time credibility assessment of content on twitter. In *International Conference on Social Informatics*. Springer, 228–243.

[26] Carolin Hahnel, Beate Eichmann, and Frank Goldhammer. 2020. Evaluation of Online Information in University Students: Development and Scaling of the Screening Instrument EVON. *Frontiers in psychology* 11 (2020).

[27] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. 2020. Array programming with NumPy. *Nature* 585, 7825 (Sept. 2020), 357–362. https://doi.org/10.1038/s41586-020-2649-2

[28] Gaole He and Ujwal Gadiraju. 2022. Walking on Eggshells: Using Analogies to Promote Appropriate Reliance in Human-AI Decision Making. In *Proceedings of the Workshop on Trust and Reliance on AI-Human Teams at the ACM Conference on Human Factors in Computing Systems (CHI '22)*.

[29] Robert R. Hoffman and Gary Klein. 2017. Explaining Explanation, Part 1: Theoretical Foundations. *IEEE Intelligent Systems* 32, 3 (2017), 68–73. https://doi.org/10.1109/MIS.2017.54

[30] Oana Inel, Tomislav Duricic, Harmanpreet Kaur, Elisabeth Lex, and Nava Tintarev. 2021. Design Implications for Explanations: A Case Study on Supporting Reflective Assessment of Potentially Misleading Videos. *Frontiers in artificial intelligence* 4 (2021).

[31] Oana Inel, Nava Tintarev, and Lora Aroyo. 2020. Eliciting User Preferences for Personalized Explanations for Video Summaries. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 98–106.

[32] Farnaz Jahanbakhsh, Amy X Zhang, Adam J Berinsky, Gordon Pennycook, David G Rand, and David R Karger. 2021. Exploring lightweight interventions at posting time to reduce the sharing of misinformation on social media. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–42.

[33] Srijan Kumar, Robert West, and Jure Leskovec. 2016. Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes. In *Proceedings of the 25th International Conference on World Wide Web* (Montréal, Québec, Canada) *(WWW '16)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 591–602. https://doi.org/10.1145/2872427.2883085

[34] Moritz Körber. 2020. Theoretical considerations and development of a questionnaire to measure trust in automation. osf.io/y3jn5

[35] David La Barbera, Kevin Roitero, Gianluca Demartini, Stefano Mizzaro, and Damiano Spina. 2020. Crowdsourcing Truthfulness: The Impact of Judgment Scale and Assessor Bias. In *ECIR'20 Proceedings of the 42nd European Conference on Information Retrieval*, Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins (Eds.). Springer International Publishing, Cham, 207–214. https://doi.org/10.1007/978-3-030-45442-5_26

[36] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and Customizable Explanations of Black Box Models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (Honolulu, HI, USA) *(AIES '19)*. Association for Computing Machinery, New York, NY, USA, 131–138. https://doi.org/10.1145/3306618.3314229

[37] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096. https://doi.org/10.1126/science.aao2998 arXiv:https://www.science.org/doi/pdf/10.1126/science.aao2998

[38] Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 107–117. https://doi.org/10.18653/v1/D16-1011

[39] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[40] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature machine intelligence* 2, 1 (2020), 56–67.

[41] John Manoogian III and Buster Benson. 2020. The cognitive bias codex—180+ biases.

[42] Brent Mittelstadt, Chris Russell, and Sandra Wachter. 2019. Explaining explanations in AI. In *Proceedings of the conference on fairness, accountability, and transparency*. 279–288.

[43] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. 2020. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. arXiv:1811.11839 [cs.HC]

[44] Preslav Nakov and Giovanni Da San Martino. 2021. *Fake News, Disinformation, Propaganda, and Media Bias.* Association for Computing Machinery, New York, NY, USA, 4862–4865. https://doi.org/10.1145/3459637.3482026

[45] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction* 27, 3-5 (Oct 2017), 393–444. https://doi.org/10.1007/s11257-017-9195-0

[46] Ziad Obermeyer and Ezekiel J Emanuel. 2016. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine* 375, 13 (2016), 1216.

[47] Heather O'Brien, Paul Cairns, and Mark Hall. 2018. A Practical Approach to Measuring User Engagement with the Refined User Engagement Scale (UES) and New UES Short Form. *International Journal of Human-Computer Studies* 112 (04 2018). https://doi.org/10.1016/j.ijhcs.2018.01.004

[48] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 8779–8788. https://doi.org/10.1109/CVPR.2018.00915

[49] Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. 2020. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science* 66, 11 (2020), 4944–4957.

[50] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2017. Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media. In *Proceedings of the 26th International Conference on World Wide Web Companion* (Perth, Australia) *(WWW '17 Companion)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1003–1012. https://doi.org/10.1145/3041021.3055133

[51] Kashyap Popat, Subhabrata Mukherjee, Jannik Strötgen, and Gerhard Weikum. 2018. CredEye: A Credibility Lens for Analyzing and Explaining Misinformation. In *Companion Proceedings of the The Web Conference 2018* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences

Steering Committee, Republic and Canton of Geneva, CHE, 155–158. https://doi.org/10.1145/3184558.3186967

[52] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. 2018. DeClarE: Debunking Fake News and False Claims using Evidence-Aware Deep Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 22–32. https://doi.org/10.18653/v1/D18-1003

[53] Kevin Roitero, Michael Soprano, Shaoyang Fan, Damiano Spina, Stefano Mizzaro, and Gianluca Demartini. 2020. Can The Crowd Identify Misinformation Objectively? *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Jul 2020). https://doi.org/10.1145/3397271.3401112

[54] Joni Salminen, Ying-Hsang Liu, Sercan Şengün, João M Santos, Soon-gyo Jung, and Bernard J Jansen. 2020. The effect of numerical and textual information on visual engagement and perceptions of AI-driven persona interfaces. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 357–368.

[55] Skipper Seabold and Josef Perktold. 2010. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*.

[56] Alison Marie Smith-Renner, Styliani Kleanthous Loizou, Jonathan Dodge, Casey Dugan, Min Kyung Lee, Brian Y Lim, Tsvi Kuflik, Advait Sarkar, Avital Shulner-Tal, and Simone Stumpf. 2021. TExSS: Transparency and Explanations in Smart Systems. In *26th International Conference on Intelligent User Interfaces - Companion* (College Station, TX, USA) *(IUI '21 Companion)*. Association for Computing Machinery, New York, NY, USA, 24–25. https://doi.org/10.1145/3397482.3450705

[57] John Sweller, Jeroen J. G. van Merrienboer, and Fred Paas. 2019. Cognitive Architecture and Instructional Design: 20Years Later. *Educational Psychology Review* 31, 2 (June 2019), 261–292. https://doi.org/10.1007/s10648-019-09465-5 11th International Cognitive Load Theory Conference (ICLTC), ICLTC2018 ; Conference date: 04-09-2018 Through 06-09-2018.

[58] John Sweller, Jeroen J. G. Van Merrienboer, and Fred Paas. 1998. Cognitive Architecture and Instructional Design. *Educational Psychology Review* 10 (09 1998), 251–. https://doi.org/10.1023/a:1022193728205

[59] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*. 109–119.

[60] Nava Tintarev and Judith Masthoff. 2007. A Survey of Explanations in Recommender Systems. *IEEE 23rd International Conference on Data Engineering Workshop*, 801–810. https://doi.org/10.1109/ICDEW.2007.4401070

[61] Nava Tintarev and Judith Masthoff. 2015. *Explaining Recommendations: Design and Evaluation.* Springer US, Boston, MA, 353–382. https://doi.org/10.1007/978-1-4899-7637-6_10

[62] Suzanne Tolmeijer, Ujwal Gadiraju, Ramya Ghantasala, Akshit Gupta, and Abraham Bernstein. 2021. Second chance for a first impression? Trust development in intelligent system interaction. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. 77–87.

[63] Thi Ngoc Trang Tran, Müslüm Atas, Alexander Felfernig, Viet Man Le, Ralph Samer, and Martin Stettinger. 2019. Towards social choice-based explanations in group recommender systems. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. 13–21.

[64] Chun-Hua Tsai and Peter Brusilovsky. 2019. Evaluating visual explanations for similarity-based recommendations: User perception and performance. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization*. 22–30.

[65] Raphael Vallat. 2018. Pingouin: statistics in Python. *Journal of Open Source Software* 3, 31 (2018), 1026. https://doi.org/10.21105/joss.01026

[66] Elmira van den Broek, Anastasia Sergeeva, and Marleen Huysman. 2020. Hiring algorithms: An ethnography of fairness in practice. In *40th international conference on information systems, ICIS 2019*. Association for Information Systems, 1–9.

[67] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2

[68] Jacky Visser, John Lawrence, and Chris Reed. 2020. Reason-Checking Fake News. *Commun. ACM* 63, 11 (oct 2020), 38–40. https://doi.org/10.1145/3397189

[69] Nguyen Vo and Kyumin Lee. 2018. The Rise of Guardians: Fact-Checking URL Recommendation to Combat Fake News. In *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR '18)*. Association for Computing Machinery, New York, NY, USA, 275–284. https://doi.org/10.1145/3209978.3210037

[70] Amber Walraven, Saskia Brand-Gruwel, and Henny PA Boshuizen. 2008. Information-problem solving: A review of problems students encounter and instructional solutions. *Computers in Human Behavior* 24, 3 (2008), 623–648.

[71] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, 422–426. https://doi.org/10.18653/v1/P17-2067

[72] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th International Conference on Intelligent User Interfaces*. 318–328.

[73] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*. 189–201.

[74] Waheeb Yaqub, Otari Kakhidze, Morgan L Brockman, Nasir Memon, and Sameer Patil. 2020. Effects of credibility indicators on social media news sharing intent. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.