# Evaluating the Impact of Human Explanation Strategies on Human-AI Visual Decision-Making

ANONYMOUS AUTHOR(S)

Artificial intelligence (AI) is increasingly being deployed in high-stakes domains, such as disaster relief and radiology, to aid practitioners during the decision making process for interpreting images. Explainable AI techniques have been developed and deployed to provide users insights into why the AI made a certain prediction. However, recent research suggests that these techniques may be confusing or misleading to users. We conduct a series of two studies to understand how different types of explanations may impact visual decision-making. In our first study, we elicit explanations from humans when assessing damaged buildings after natural disasters from satellite imagery and identify four core explanation strategies that humans employed. In addition, we study the impact of these explanation strategies with a different set of decision-makers performing the same task. We provide initial insights on how causal explanation strategies improve humans' accuracy and calibrate humans' reliance on AI when the AI is incorrect. However, we also find that causal explanation strategies may lead to incorrect rationalizations when the AI presents a correct assessment with incorrect localization.

CCS Concepts: • **Human-centered computing** → **Collaborative and social computing**; • **Applied computing** → **Computers in other domains**.

Additional Key Words and Phrases: Explanation Generation, Human-Centered Explainable AI, Human-AI Collaboration

**ACM Reference Format:**
Anonymous Author(s). 2018. Evaluating the Impact of Human Explanation Strategies on Human-AI Visual Decision-Making. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 33 pages. https://doi.org/10.1145/1122445.1122456

## 1 INTRODUCTION

Artificial intelligence (AI) systems are increasingly being deployed with the aim of helping practitioners make high-stakes decisions more quickly and accurately. Computer Vision for deep learning focuses on understanding our world visually by classifying and segmenting images among many other tasks [57]. With computer vision models rapidly increasing in accuracy, several tasks requiring image classification, segmentation, or object detection have been automated to aid practitioners. For example, radiologists work with AI-based tools to evaluate medical imagery [8, 46], while GIS (geographic information system) experts work with AI to assess building damage from satellite imagery after natural disasters [18]. Human-AI collaboration is increasingly prevalent across a range of high-stakes decision-making contexts. However, human-AI collaboration may fail to result in better decision-making in practice due to misleading explanations or explanations that were designed with the intuition of a different stakeholder. This can lead to over- or under-reliance on the AI resulting in costly misinterpretations and misuse.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
*Woodstock '18, June 03–05, 2018, Woodstock, NY*
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

To date, several explainable AI (XAI) techniques have been developed with the goals of providing insight into how an AI makes its decisions, helping decision-makers calibrate their trust and reliance upon AI[1], and improve human-AI collaboration overall. For instance, some techniques yield saliency maps (heatmaps) indicating the most important regions of the image that contributed to its classification. However, to date, many of these methods have been designed for data scientists or machine learning engineers instead of subject matter experts [34]. Furthermore, Brennen [6] and Liao et al. [34] emphasize that several XAI techniques do not present the information that decision makers actually need to see to make accountable, accurate decisions. For instance, in the context of medical imaging, Saporta et al. [51] report that existing XAI techniques rarely highlight clinically meaningful image regions. These studies highlight the need for a human-centered approach to designing and evaluating AI explanation techniques, to understand which explanation techniques, or combination of techniques, can truly support effective human-AI collaborative decision-making.

To inform the design of AI explanations, there is much to learn from *human-generated* explanations in the context of human-human collaboration. In many real-world decision-making settings, practitioners generate explanations intended for other practitioners to consume, to support their decision-making. For example, when radiologists make observations from medical images, they often characterize their observations and inferences for other radiologists or practitioners in other care units [9].

Inspired by these observations, in this study, we focus on three primary questions:

- **RQ1.** What strategies do humans use to explain, or rationalize, their reasoning for a high-stakes image classification task?
- **RQ2.** How do different human explanation strategies improve humans' accuracy in a high-stakes image classification task?
- **RQ3.** How do different human explanation strategies impact decision makers' reliance on AI?

We focus on the complex task of assessing building damage from satellite imagery after natural disasters. This task is important to the Humanitarian Assistance and Disaster Relief (HADR) community, but research designing XAI techniques and studying human-AI collaboration in this context has been scarce.

We address our research questions through a series of two studies. For the first study, we designed a system that shows participants several pairs of images for a specific geographic location before and after a natural disaster. For each image pair, participants assessed the level of damage and provided explanations and annotations for their assessment. The results allowed us to identify and analyze explanation strategies humans use when assessing image pairs.

In the second study, a new set of participants performed the same damage assessment task but were also shown "AI-based" recommendations and explanations to help them make their damage assessment. While these explanations were presented to participants as if they were generated by an AI system, they were actually human-generated explanations collected from our first study. This approach allowed us to evaluate how different explanation strategies identified from the first study may impact the decision makers' accuracy as well as their reliance on AI recommendations.

Across our two studies, we contribute the following:

- We provide insight into the **kinds of explanation strategies humans employ** when providing rationales for their decisions, in the context of visual decision-making tasks.

[1]Throughout this paper, we distinguish between the terms *reliance* and *trust*: reliance is an observable human behavior that can be measured whereas human trust is a latent factor, which cannot directly be measured from observable human behaviors [11, 22, 33, 59]

- We present a **new approach** for exploring the impacts of prospective explainable AI techniques on human-AI decision-making, by presenting participants with different types of human-generated explanations, framed as AI explanations.
- Using this approach, we conduct the first empirical investigation in the literature into the **impacts of different human explanation strategies** on human accuracy and reliance upon AI-based assessments.

## 2 RELATED WORK

As AI is supporting more decision-making processes, providing explanations to decision makers on how AI makes predictions is becoming increasingly important. On such an account, several explainable AI techniques have been proposed to provide insight into what features contributed to the predictions from AI. We present several explainability techniques from XAI literature, as well user studies on different XAI techniques from human-computer interaction literature.

### 2.1 Explainable AI Techniques

Explainable AI (XAI) is increasingly being developed for a wide range of tasks, from helping data scientists debug machine learning models to aiding doctors while diagnosing patients. For some tasks, the model being used is inherently interpretable (i.e., interpretable models); for other tasks, interpretability is completely lost and requires an additional method to understand the prediction or mechanics inner-workings of the model [42].

Specifically, computer vision tasks such as image classification or object detection employ deep neural networks (DNNs), or black-box models, that are quite difficult to interpret without XAI techniques. To address this issue, several model-agnostic techniques have been devised (techniques that can be used regardless of the model) to provide insight into the abstractions of these models [50]. These model-agnostic techniques are either providing insights to the overall limitations and capabilities of the model (global explanations) or offering insights into individual predictions (local explanations) [1, 42].

Adadi and Berrada [1] provide an in-depth survey of several different types of explainability methods. We briefly review some of the techniques that are most prevalent throughout the Human-AI collaboration and empirical XAI literature.

LIME and SHAP are often considered two of the most popular and prominent local, model-agnostic techniques [36, 56]. For instance, LIME [49], is a local, model-agnostic technique that uses local linear approximation for explaining outputs. For image data, it shows grouped regions (i.e., superpixels) of an image to highlight the most important feature that contributed to the classification of the image [50]. SHAP [39], shows the importance of each feature for one prediction [38]. In addition to these techniques, other local, model-agnostic techniques such as GradCAM and XRAI are devised to show the saliency maps, or heatmaps, of the most salient, or important, region of the image that contributed to its classification [24, 52, 55]. While GradCAM and XRAI show feature attributions by highlighting regions of the image, feature visualizations provide an insight into what the model has learned by generating examples [44].

Another technique called example-based explanations provides insight into the AI's limitations and capabilities on a task by identifying certain instances that showcase such limitations and capabilities [1]. One specific type of example-based explanations is a counterfactual explanation which takes a given prediction and provides a statement about what the prediction would have been if a certain feature had a different value [60]. This explanation technique has recently become popular within the visual decision-making for image classification [17, 32].

In line with the feasibility of XAI techniques and their technological advances, AI-assisted high-stakes decision-making processes are also getting attention in adopting XAI. As such, researchers

have attempted to attach explanations in AI-powered high-stakes situations, such as decision-making in humanitarian assistance and disaster relief (HADR). For instance, previous studies presented the use of the SHAP technique for explaining the results of HADR detection models trained for various tasks, such as earthquake-induced building damage detection [40] and spatial drought prediction [12]. In addition to the deployment of XAI techniques, Andres et al. provided insights on forecasting several application methods of in supporting the decision-making processes of humanitarian aid planners with the aid of XAI [2].

### 2.2 Human-Centered Explainable AI Techniques

The majority of XAI work has focused on *interpretability* instead of *explanation generation* which Ehsan and Riedl [14] define as providing, *"... useful information for practitioners and users in an accessible manner."* In this work, they proposed taking a sociotechnical approach to XAI given the dynamic situations between humans and XAI systems. For example, Ehsan et al. [13] trained a DNN on data from humans speaking aloud while playing a game to train an AI to rationalize how it plays that game. Their user study showed that players were more satisfied with the generated rationalizations than other explanation methods [13].

Hendricks et al. [19] proposed a model that generates justifications, or visual explanations, for image classification by including class discriminative features to provide specific distinguishable information in hopes to aid non-experts. However, the explanations are not generated based on data from human explanations; rather, they are generated based on visual features and fine-grained visual descriptions [19].

While explanations generated from natural language techniques are more intuitive to end-users, Sevastjanova et al. [53] emphasized the importance of combining multiple types of explanations (i.e., visualizations, text) to generate explanations for machine learning models. They also provided examples of what an explanation that combines visualizations and text might look like.

Several explainable AI techniques have been designed to provide insight into "black-box" models; however, those techniques are not informed by the types of explanations humans generate. Until recently, XAI techniques were not evaluated with end-users to confirm which types of explanations end-users find helpful, what information the end-users are looking for, or who the end-users are. Explanations can be requested for a variety of reasons from a variety of different end-users; incorporating *who* the explanation is being designed for and *why* is a core part of human-centered XAI [15]. Furthermore, human-centered XAI is interdisciplinary combining cognitive science, design, and sociotechnical perspectives [35].

### 2.3 Impact of Explainable AI on Humans

Previous studies have evaluated the impact of different XAI techniques on humans, such as trust, reliance, and task-performing accuracy. For instance, Zhang et al. compared the impact that local explanations and model confidence have on the human's trust in AI, particularly when collaborating on a task where the human and AI have similar performance [64]. As a result, they found that participants' trust in the model increased when the AI's confidence in its prediction was high. They did not observe any impact from local explanations being shown to the participants [64]. Similarly, Bansal et al. [4] evaluate the impact of saliency explanations and expert-generated explanations on sentiment analysis and question answering tasks where the human and AI have similar performance. They observed that humans were more likely to agree with the AI when shown explanations even when the AI was incorrect [4].

Chu et al. [10] explore the impact of saliency maps on an age prediction task when the saliency maps highlight meaningful regions, spurious regions, and randomly generated regions of the image. They did not find that the saliency maps improved the participants' accuracy or trust. Nourani

et al. [43] conducted a similar experiment evaluating the impact of meaningful and meaningless saliency maps on the perception of system accuracy. They observed that meaningless saliency maps negatively impacted how the participants perceived the system's accuracy.

Wang and Yin [62] compare the impacts of four different explanation techniques including counterfactual explanations, feature importance, feature contribution, and nearest neighbors on two different tasks. Overall, their results find that counterfactuals did not help calibrate trust [62]. Similarly, another study designed an interactive system for predicting the risk of child maltreatment with four different explanation techniques [65]. They recruited experts and non-experts for their study and observed that feature contribution was the most useful explanation across experts and non-experts.

As a result of these user studies, it is important to understand how humans generate explanations to improve current explanation techniques and human-AI decision-making. To our knowledge, there is no study that identifies defined strategies from human explanations and identifies their effects on human-AI decision-making through in-depth qualitative and empirical study.

Thus, in this study, we extend the literature by identifying how humans explain in a visual decision-making task and characterize their explanations into core explanation strategies. With these strategies, we further evaluate the effects (i.e., assessment accuracy and reliance on AI assessment) of each strategy on humans.

## 3 TASK SELECTION: BUILDING DAMAGE ASSESSMENT

Satellite imagery is abundant which has resulted in numerous computer vision applications. For example, satellite imagery and computer vision techniques have been used in various high-stakes scenarios, such as identifying economic growth and stability [23, 45, 63], detecting poachers and illegal fishing vessels [58], and identifying damaged buildings after natural disasters [18].

Natural disasters, such as hurricanes, flooding, wildfires, and earthquakes often have devastating effects on humans and their properties. Assessing building damage after a natural disaster from satellite imagery is a critical task. The damaged structures put a huge strain on the local and regional economies forcing officials to rely on the government for funds to support recovery efforts. Our interviews with HADR experts suggest that in certain situations to receive funding in a timely manner for recovery efforts, rapid and accurate identification of the number of damaged is crucial.

### 3.1 Task Expertise

We chose building damage assessment from satellite imagery as our task because prior work indicates that it is possible to quickly train users to perform this task with reasonable accuracy [26]. Unlike other specialized domains such as medical imaging, where users are often required to have rare expertise, prior work has successfully used crowdsourcing approaches to help gather information more quickly after natural disasters [26, 54]. Furthermore, many people have experience viewing and interpreting satellite images in their day-to-day life, such as when using navigation systems or online interfaces like Google Earth.

### 3.2 Task Generalization

Detecting damaged structures from satellite imagery is analogous to several computer vision tasks. Building damage assessment in satellite imagery requires the analysts to pan and zoom in on a pair of before and after images to identify any abnormalities. The task of comparing two images to identify and annotate differences between images generalizes beyond damage assessment. For example, radiologists compare multiple views to make diagnostic interpretations in breast imaging[16]. Thus, we believe that insights derived from satellite imagery assessment tasks could be extensible to other domains involving computer vision applications.

### 3.3 Selected Dataset

For training participants to assess building damage from satellite imagery, we utilized satellite imagery data from the xBD dataset [18]. Designed for supporting humanitarian assistance and disaster relief research, the xBD dataset covers a wide range of natural disasters that frequently occur, such as floods, earthquakes, wildfires, and hurricanes. This dataset offers before and after images as well as a ground truth assessment and annotation for the damage. With the ground truth available, we were able to provide training for participants to get familiar with the task at hand.

Since most participants are likely to be new in assessing building damage from satellite imagery, we deemed it not beneficial to present participants images that had more than one damaged structure with different levels of damage. This choice was made also to reduce potential confounders. We filtered out 96.53% of images from the xBD dataset that had more than one damaged structure with different levels of damage or no damaged structures.

## 4 STUDY 1: HUMAN EXPLANATION STRATEGIES

In this study, we recruited online participants to elicit human explanations during visual decision-making. The main goals of this study are to understand:

- What strategies emerge when they describe the classification of a pair of images?
- What are the most prevalent strategies among all of the strategies that are identified?

### 4.1 Study Design

*4.1.1 Recruitment & Participants.* To collect explanations, we performed our study on an online participant recruiting platform, Prolific. For a consistent participant base, we recruited workers who use English as their first language, currently reside in the United States, have an approval rate of 95%, and have a record of at least 50 task submissions. After screening out incomplete participants, responses from a total of 60 workers ($M_{age}$ = 34.07, $SD_{age}$ = 8.28, 26 female) were collected. Once completing the procedure, each worker was assigned an anonymous ID and compensated 6.50 USD each for their participation. In order to motivate participants, we offered a bonus payment of 1 USD to those who received the highest scores on their assessments (25% of participants were awarded bonuses).

*4.1.2 Data Selection Method.* About half of the image sets in the xBD dataset contain no building damage (46.37%). Furthermore, of the images with damaged buildings, 89.47% contained multiple buildings. Our experimental design focuses only on images with either one damaged building (5.64% of images) or multiple damaged buildings all with the same damage level (18.26%) so we could isolate the assessment and localization for one individual target. Furthermore, this allowed us to focus our training for a single target. For the training session, we curated a set of 27 image sets for a variety of natural disasters (7 flooding, 13 hurricane, 5 wildfire, 2 tsunami). For the main session, we curated a set of 10 cases for a variety of natural disasters (4 hurricane, 3 flooding, 2 wildfire, and 1 earthquake), each of which contains pre- and post-disaster images with a single damaged building. The chosen images we used are available on our temporarily anonymous GitHub supplement: training set, main set.

*4.1.3 Training session.* Before getting into the main session, participants were required to complete a training session. The objective of the training session is to let participants ease into the annotation interface and become familiarized with building damage assessment tasks. Participants are provided detailed criteria for assessing damages on the interface to help learn the differences between the levels of damage [18]. With these criteria in mind, participants are asked to complete the training session by assessing damage in several sets of satellite images. The training interface seen in

Figure 1 shows the pre- and post-disaster satellite images with four different damage assessment options.

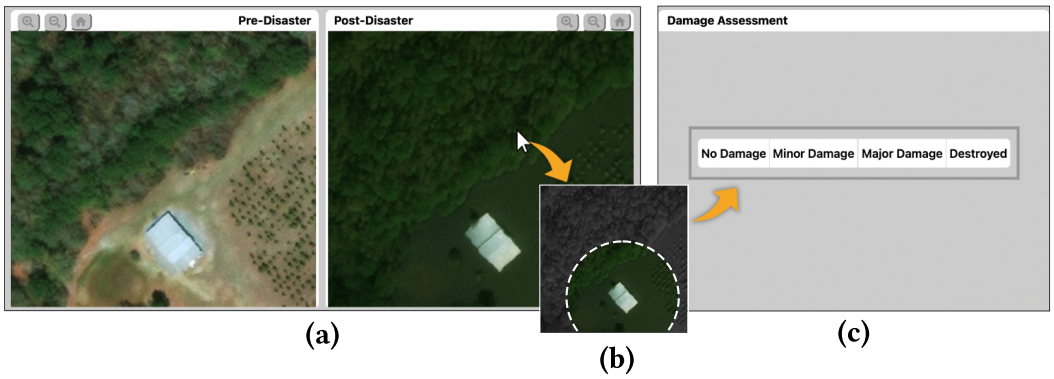



Fig. 1. Interface of the training session used to familiarize users with the task of building damage assessment. (a) Two satellite images side-by-side are presented, and (c) the user assesses damage based on four different options: No damage, minor damage, major damage, and destroyed. For the first and second sub-sessions, (b) users are shown an overlay in the image highlighting where the damage is when hovering to help them locate the building

The training phase is composed of three sub-sessions with slightly different functionality in each to help familiarize participants with the task and navigate the interface. Each sub-session shows nine pairs of satellite images in groups of three. The participant is alerted which damage assessments they got incorrect and correct. The participant must select the correct damage assessment for each pair of images in a group before seeing the next group. The average time each participant spent completing training tasks was 13.26 minutes ($SD$ = 5.52).

The three sub-sessions are slightly different from one another. The first sub-session shows a clue (seen in Figure 1-(b)) of where they should look to assess the level of damage when they hover over the imagery. However, they do not have zoom or pan functionalities. The second sub-session adds in zoom/pan functionality with the clue. Finally, in the third sub-session, the clue is taken away, but the zoom/pan functionality remains. The sub-sessions were designed to help the participant understand what features to look for in satellite images when assessing the damage.

Even if the participants may have become sufficiently accustomed to the interface and assessing damages in the training session, we presumed that they still may not be familiar with generating annotations on top of the images. Thus, we let participants go through a tour of the annotation interface designed for the main session. This tour introduces the participants to the different annotation tools offered and how to add text to rationalize their annotations.

*4.1.4 Main session.* After successfully completing the training session, each participant was presented with a pair of satellite images of a specific geographical region before and after a natural disaster. We presented 10 sets of satellite images that we curated from the xBD data set [18]. Participants were asked to assess the damage in all 10 image sets. To mitigate possible biases stemming from the sequence, we assigned each participant to one of four different sequences for presenting imagery pairs.

While assessing image sets, participants were offered a wide range of drawing tools (rectangle, polyline, pin, circle, pen) and viewing functionalities (pan, zoom, mirror-drawing), as shown in Figure 2. Mirror-drawing allows participants to synchronize their annotations for the pre- and

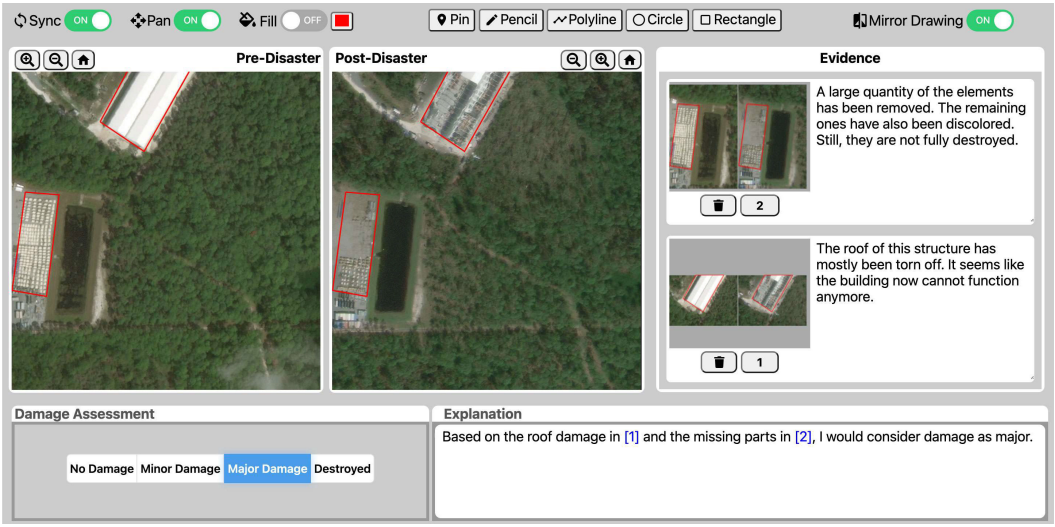



Fig. 2. Overview of the interface for the main task. Participants are presented with several different annotation tools to highlight evidence that helps explain their damage assessment. Participants provide explanations for each annotation they add as well as a global explanation for the pair of images overall

post-disaster images. When mirror-drawing is selected, the participant only has the draw their annotation once and it will be reflected on the other image in the same location. The annotation tools allowed participants to have flexibility in locating damaged buildings and choosing how to generate visual components of their explanations.

Once participants draw a particular markup using the provided drawing and viewing tools, our system shows a text field to allow participants to explain the marked-up region via text. The marked-up region is noted as the image-based annotation and the accompanying text is noted as the text-based annotation. Participants can make references to the image- and text-based annotations by citing them within their global explanation (seen in Figure 2).

After they successfully completed the task, each participant was then asked to assess the clarity of the guidelines we offered, along with their own performance while completing tasks based on the 5-point Likert scale. Additionally, in order to identify possible enhancements for designing the interface, we asked them to freely note any tools that were difficult to make use of and the tasks that were particularly difficult, as well as their comments on the study. The average time elapsed for the main phase was 23.71 minutes ($SD$ = 12.19). The average number of image- and text-based annotations per participant was 14.73 annotations ($SD$ = 4.07) and an average of 1.47 annotations per image ($SD$ = 0.85).

*4.1.5 Analysis.* We used thematic analysis [5] to identify insights about how people generate explanations for their damage assessment. Specifically, our goal was to extract and categorize the emerging themes that may illustrate how each participant rationalizes their damage assessments. For our analysis, two coders independently analyzed the responses (image- and text-based annotations) and made notes on how each participant made explanations. Once complete, coders set up initial themes and grouped each response into one or multiple themes. Until making a final consensus on every response, coders iterated over the response groupings to modify and finalize each theme.

Table 1. Summary of the strategies we identified during the study. Grey rows indicate major codes, each of which is followed by the sub-codes (except for the code C which contains no sub-codes)

| Code | Definition | Example |
|---|---|---|
| A | Constructing a causal story to explain structural damage | |
| A-1 | Use evidence of natural disaster/lower-level cause to argue that there was damage on objects | *"From the evidence of flooding, the structure seems to have been affected"* |
| A-2 | Use evidence of structural / sub-structural damage / surrounding area to argue that there was the effect of natural disaster/lower-level cause | *"Since the structure seems to be underwater, there must have been a flood"* |
| A-3 | Multi-step causal chain argument | *"From the evidence of the flood, I suspect the flood has moved and damaged the structure"* |
| B | Contrasting pre- and post-disaster imagery | |
| B-1 | Comparison of the structure | See Figure 4-(a) |
| B-2 | Comparison of the surrounding area | See Figure 4-(b) |
| B-3 | Comparison of the sub-structures | See Figure 4-(c) |
| B-U | Falls under B, but without using a mirror-drawing or objects unidentifiable | See Figure 4-(d) |
| C | Highlighting affected part of a building | See Figure 4-(c) |
| D | Explanations based on the extent of damage to a specific building | |
| D-1 | Based on the amount or proportion of the structure that appears to be damaged | *"The building is not completely destroyed but it has lost the majority of its roof"* |
| D-2 | Based on their inference about the possibility of being recovered someday | *"Looks like it can be recovered someday"* |
| E | Explaining reasons for lack of confidence in their own assessment | |
| E-1 | Due to confusing artifacts in images | *"It's hard to tell due to the shadow"* |
| E-2 | Due to the changes irrelevant to disaster between pre- and post-, or the elapsed time between them | *"Seems like new buildings had been built"* |
| E-U | Other reasons or without any reason | *"This image is a bit mystifying to decode"* |
| F | Using the number of damaged structures in an image as the measure for severity of the disaster | |
| F-1 | Structures only | *"There are some structures remaining but most have been very damaged"* |
| F-2 | Structure + surrounding area | *"Small area burnt. Home intact"* |
| F-3 | Surrounding area only | *"Majority of the water has been removed"* |
| F-U | Non-identifiable | Simply indicating region as 'area' |
| O | Other minor codes | |
| N | Simply noting as 'No damage' | |

## 4.2 Results

*4.2.1 Summary of the Identified Strategies.* Six major strategies were elicited, where we gained the high overall inter-rater reliability (Krippendorff's $\alpha$) of 0.75. The definition and the concrete example for each strategy are listed in Table 1, and the frequency of each strategy was plotted in Figure 9 in Appendix Section A.2.

*4.2.2 Representative Strategies.* We provide definitions and examples of the six major strategies we identified (Table 1). Although we identified six major strategies (A - F), strategies C and E were so sparse (about 10% of overall user assessments) that it was impossible to identify the exemplar case for certain pairs of satellite images. Thus, strategies A, B, D, and F were curated as the *core explanation strategies*. In this section, we further describe the detailed criteria and explanation of the most prevalent strategies.

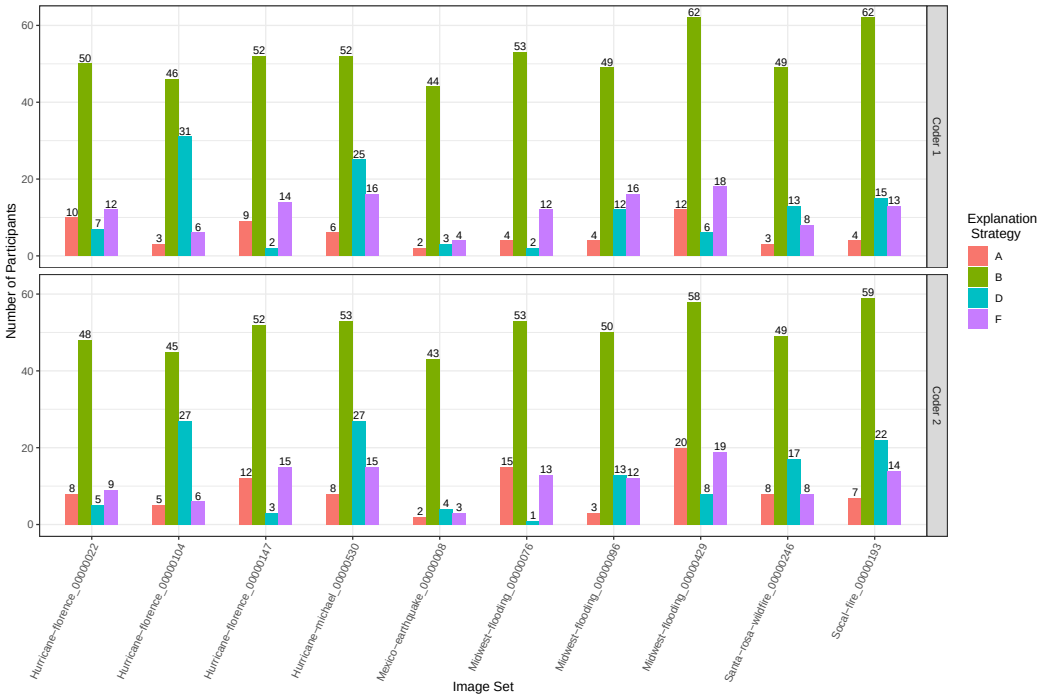
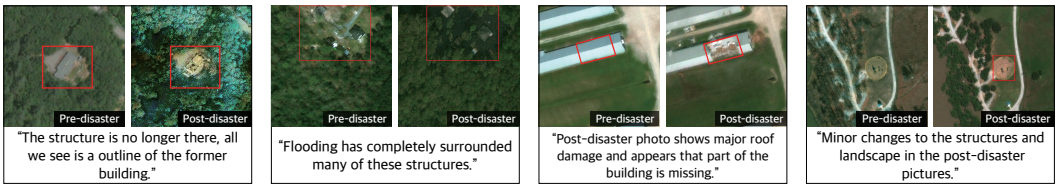



Fig. 3. The number of participants that used each core explanation strategy for each image set was reported by both coders. Some participants used multiple explanation strategies in their response to a given image set.

***Strategy A: Constructing a causal argument to explain building damage.*** Rather than directly referencing visual features of a building itself, participants instead pointed to visual evidence of a natural disaster in a building's surroundings to explain their assessment of building damage (A-1; e.g., *"From the evidence of flooding, I would say the building seems to have been affected"*). In other cases, participants inferred that a particular type of a natural disaster had occurred based on evidence of damage to a building, and then explained their overall assessment of building damage with reference to the type of disaster (A-2; e.g., *"The building has roof damage. Probably a hurricane came and hit it"*). Some participants constructed more complex, multi-step causal arguments (A-3;




**(a) Example of B-1** **(b) Example of B-2** **(c) Example of B-3 and C** **(d) Example of B-U**

Fig. 4. Examples of sub-strategies B-1, B-2, B-3, B-U, and C

e.g., *"**(Step 1)** There was a fire and **(Step 2)** it was a wildfire that took everything from the building. **(Step 3)** You can only see the outline of the building"*).

***Strategy B: Contrasting pre- and post-disaster imagery.*** As seen in Figure 3, a majority of participants explained their damage assessments through direct comparisons across image sets (e.g., by creating an annotation using the mirror-drawing tool). In more than 200 cases, participants referenced contrasts in the appearance of a specific building between the pre- and post-disaster images (B-1; see Figure 4-(a)), or contrasts in the appearance of specific sub-structures of a building (B-3; see Figure 4-(c)). In addition, participants often directly compared the pre- and post-disaster appearance of the area surrounding a building (B-2; see Figure 4-(b)). Finally, ambiguous cases in which people generated contrast-based explanations, but did not clearly specify which elements of an image they were comparing, were marked with the sub-strategy B-U (see Figure 4-(d)).

***Strategy D: Explanations based on the extent of damage to a specific building.*** For some cases, participants explained their assessment of the level of damage to a given building based on the *proportion* of the building that appears to be damaged (D-1; e.g., *"Approximately a half of the building was collapsed"*). Interestingly, even when significant building damage was evident, some participants explained lower assessments of damage by arguing that the damage appeared repairable (D-2; e.g., *"One part (of the building) was hit ... seems like it could be rebuilt"*).

***Strategy F: Using the number of damaged structures in an image as the measure for severity of the disaster.*** Whereas strategy D captures cases where participants explain their damage assessments with reference to the apparent extent of damage to the building itself, strategy F is when participants explain their damage assessments with reference to the overall extent of damage observed in the image as a whole. For example, some participants explained their building damage assessment with reference to the number of other buildings that appeared to be affected (F-1; e.g., *"It appears that one building has disappeared, leading me to believe it was destroyed. However, the remaining buildings seen are unharmed"*).

Furthermore, other participants explained their damage assessment with reference to the extent of damage visible in the area surrounding a building, either including building damage (F-2; e.g., *"None of the large buildings appear to be damaged, but there is evidence of a large mud patch (in the surrounding area), indicating some minor flood damage"*) or excluding building damage (F-3; e.g., *"All trees have been damaged or destroyed"* and *"Flooding has completely surrounded many of these structures"*). Finally, ambiguous cases were marked as F-U (e.g., *"Every area was totally destroyed"*).

## 5 STUDY 2: IMPACT OF EXPLANATION STRATEGIES

Based on our results from Study 1, we conducted a follow-up study to explore whether and how each explanation strategy impacts decision-makers' performance.

### 5.1 Study Design

In this study, participants were shown human-generated annotations and damage assessments from the first study, which were presented as if they were assessments from an "AI system". This setup allowed us to evaluate whether and how particular human explanation strategies impact decision-makers' accuracy and reliance on AI.

We modified the web-based system from the first study to show conditions consecutively to the user as they assess building damage for a given pair of satellite images, as shown in Figure 5. The code for our web-based system is available as an open-source repository (link hidden for anonymity).

*5.1.1 Recruitment & Participants.* We followed the same recruitment procedures as in Study 1: we collected results from 60 participants ($M_{age}$ = 34.93, $SD_{age}$ = 12.51; 33 female) through Prolific, who have English as their first language, who currently reside in the United States, who have an approval rate of at least 95%, and who have a minimum of 50 submissions. We excluded workers who participated in Study 1 due to the significant overlap of content. We also excluded two participants who failed an attention check. Each worker was compensated $6.50 USD each for their participation, with an anonymous ID assigned for analyzing their responses. In order to incentivize active, high-quality participation, we offered a bonus of $1 USD for those who scored within the top 50% of all participants. Out of the 60 participants, 30 participants received bonuses.

*5.1.2 Data Selection Method.* The satellite images and explanations presented to participants in this study were selected from the responses from the first study. The curated set is available on our temporarily anonymous GitHub site. One coder analyzed every observation from the first study to evaluate the quality of the visual and text-based annotations provided. The coder reviewed observations where the participant's damage assessment was correct, where both coders from Study 1 agreed on the explanation strategy, and where the explanation strategy was a core explanation strategy from Study 1. From the filtered observations, the coder mapped the visual annotations into three categories:

- **Correct Localization**: The participant was able to identify the ground-truth position of the damaged building with a visual annotation.
- **Partially Correct Localization**: The participant's visual annotation included a portion of the ground-truth position of the damaged building.
- **Incorrect Localization**: The participant's visual annotation does not include any portion of the ground truth position of the damaged building.

After coding the annotations in each observation, the coder evaluated the quality of the explanations for observations where the participant's damage assessment was correct. We specifically selected observations that only used one explanation strategy to allow us to isolate the effectiveness of each strategy by itself. We chose from observations that did not cite the annotations[2] (66% of observations and 63% of participants did not cite annotations[3]) to limit potential confusion in references and separate the impact of the text-based annotations from the global explanations.

*5.1.3 Training.* The training for this study consisted of two phases: the first phase was the same training phase from the first study (seen in Figure 1) where participants are shown nine pairs of satellite images in groups of three to get familiarized with building damage assessment. In the second training phase, we provide a walk-through of the damage assessment tool and highlight the additional information that is provided in each stage. Each participant is assigned a pair of satellite

[2] As seen in the explanation in Figure 2

[3] See Figure 11 in Appendix A.2 for full results

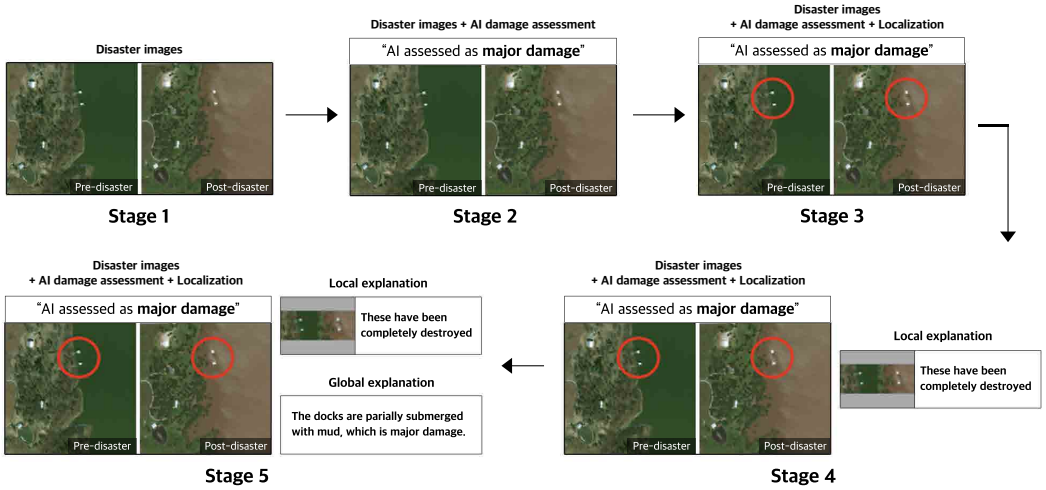



Fig. 5. The five stages are shown while completing the task for Study 2. **Stage 1** shows the pre- and post-disaster image, with no information from the AI; **Stage 2** adds information by showing how the AI assessed the damage; **Stage 3** adds visual annotations, highlighting points or regions of the image where the AI thinks it is important to look; **Stage 4** adds local text-based annotations to elaborate on each visual annotation; and **Stage 5** adds a global explanation detailing why the AI made the damage assessment that it did

images based on the satellite images they will see during the main task. The visual and text-based annotations for the example task were manually created (Appendix B.1).

*5.1.4 Main Task.* Participants were asked to assess building damage from several pairs of pre- and post-disaster satellite images. During the task, participants in this study were presented with human-generated outputs (framed as AI outputs) sampled from four different scenarios that arose in data from the first study:

- **Correct assessment with correct localization**: The Study 1 participant's assessment and visual annotation matched the ground truth assessment and ground truth position of the damaged building.
- **Correct assessment with partially correct localization**: The Study 1 participant's assessment matched the ground truth assessment while their visual annotation included a portion of the ground truth position of the damaged building.
- **Correct assessment with incorrect localization**: The Study 1 participant's assessment matched the ground truth assessment while their visual annotation was completely wrong.
- **Incorrect assessment**: The Study 1 participant's assessment was incorrect.

Before recommending that researchers start developing techniques to create explanations in a similar manner to humans as we found above, we need to evaluate whether they might actually be helpful. In this study, we presented human-generated explanations as coming from an AI system in order to understand what effects such explanation strategies might have if they were automated. We identified a set of representative examples for each explanation strategy and each scenario detailed in Section 5.1.2. Participants were randomly assigned to one of four groups (Table 2) with a total of 15 participants for each group. Each participant saw a total of four different image sets, each corresponding to one of the four scenarios. Different explanation strategies were paired with

Table 2. Four different groups were created to evaluate each explanation code in each scenario. We show which code was assigned to which scenario for each group and which image set was used for that explanation code/scenario.

| Group # | Scenario | Explanation Strategy | Image Set |
|---|---|---|---|
| 1 | *Correct Assessment, Correct Localization* | A | *Hurricane Florence 22* |
| | *Correct Assessment, Partially Correct Localization* | B | *Hurricane Michael 530* |
| | *Correct Assessment, Incorrect Localization* | D | *Midwest Flooding 96* |
| | *Incorrect Assessment* | F | *Midwest Flooding 76* |
| 2 | *Correct Assessment, Correct Localization* | B | *Socal Fire 193* |
| | *Correct Assessment, Partially Correct Localization* | D | *Hurricane Florence 22* |
| | *Correct Assessment, Incorrect Localization* | F | *Hurricane Michael 530* |
| | *Incorrect Assessment* | A | *Santa Rosa Wildfire 246* |
| 3 | *Correct Assessment, Correct Localization* | D | *Hurricane Michael 530* |
| | *Correct Assessment, Partially Correct Localization* | F | *Midwest Flooding 96* |
| | *Correct Assessment, Incorrect Localization* | A | *Hurricane Florence 22* |
| | *Incorrect Assessment* | B | *Hurricane Florence 147* |
| 4 | *Correct Assessment, Correct Localization* | F | *Hurricane Florence 104* |
| | *Correct Assessment, Partially Correct Localization* | A | *Hurricane Florence 22* |
| | *Correct Assessment, Incorrect Localization* | B | *Mexico Earthquake 8* |
| | *Incorrect Assessment* | D | *Midwest Flooding 429* |

different scenarios across the four groups. In addition, within each group, these image sets were presented to the participant in a random order to minimize possible order effects.

Within each scenario, the user went through five stages as shown in Figure 5. In each stage, the participant was asked to provide their damage assessment and use a pinpoint marker to identify exactly where they thought the damage was in the image. For each image, we also asked the user to indicate how confident they felt in their damage assessment (on a 4-point Likert scale from no damage to destroyed) and how helpful they thought the added information was (on a 4-point Likert scale from very unhelpful to very helpful) when making their damage assessment. Finally, we provided an optional text box to allow the user to briefly detail whether and how they believed the AI assessment and explanations may have affected their own damage assessment. The task took on average 28.79 minutes ($SD_{time}$ = 13.37 minutes).

## 5.2 Results

We evaluated the core explanation strategies to identify whether and how these may have impacted participants' assessment accuracy and reliance on AI assessments. To address **RQ2**, we present participants' accuracy across the five stages for each explanation strategy and scenario in Figure 6. Assessment accuracy is calculated by the percentage of participants that selected the ground truth assessment of the damage. Localization accuracy is calculated by the percentage of participants that placed the pinpoint on the ground truth damaged area. Stage 1 shows participants' baseline accuracy on the images presented in a given scenario *before* they are shown any information from the AI. Stage 2 shows participants' accuracy after they are presented with the AI's damage assessment (but before they can see the AI's annotations or explanation). Throughout the remainder of this section, we evaluate the impacts of a given *explanation strategy* by examining the change in a given

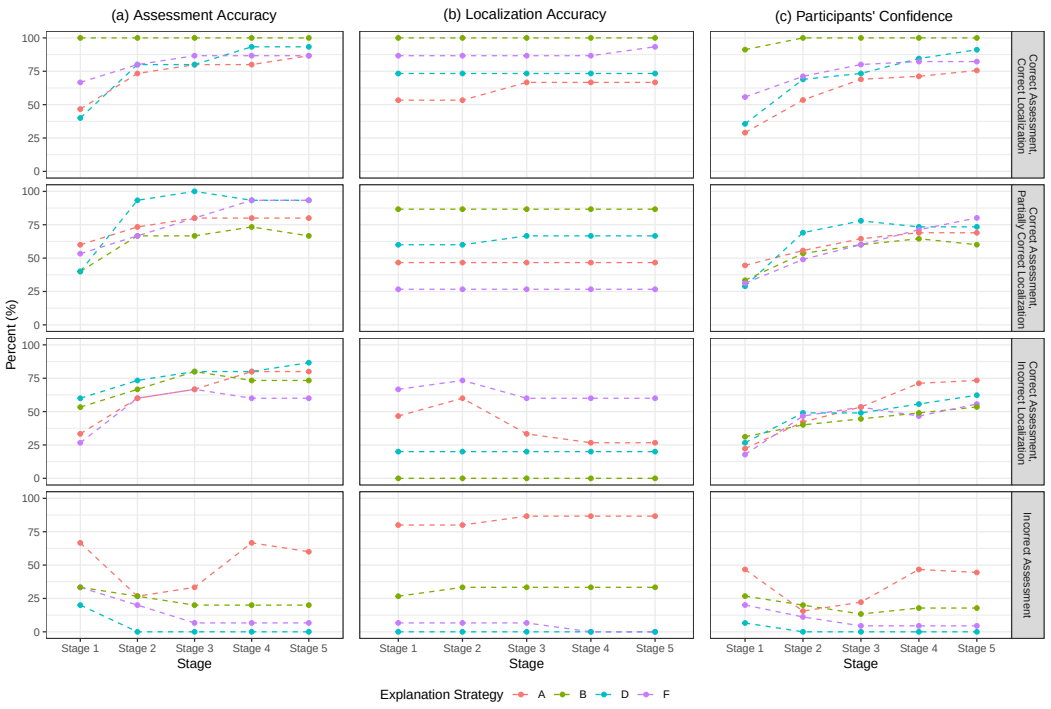



Fig. 6. Assessment accuracy, localization accuracy, and the participants' confidence. The facet on the y-axis shows the four different scenarios that the participants were presented. The percent on the y-axis represents the ratio of participants (a) that got the damage assessment correct, (b) that placed the pinpoint on the damaged structure, and (c) the confidence level for correct assessments. The x-axis shows the five stages described in Figure 5

measure between Stage 5 (i.e., all AI outputs shown) and Stage 2 (i.e., AI damage assessment only). Where appropriate, we also examine changes between other stages, for example, to understand the impacts of presenting text-based annotations over and above visual annotations. We provide a detailed interpretation of our results below (see Table 6 in Appendix Section B.3 for full results).

*5.2.1 Causal explanation strategies mislead humans less.* We found that, in scenarios where the AI damage assessment was actually incorrect, causal explanations (strategy A) misled humans less than other explanation strategies. An ANOVA for the *incorrect assessment* scenario ($p < 0.001$), indicated a significant difference in the impacts that different explanation strategies have on assessment accuracy. As shown in Table 3b, a post-hoc Tukey HSD test for this scenario revealed a statistically significant difference in the Stage 5 - Stage 2 slope between explanation strategy A (involving a causal argument to explain a building damage assessment) and all other core strategies. This may suggest that humans are better at calibrating their reliance on AI assessments by reasoning about a causal argument, versus by assessing other kinds of explanations. In line with this interpretation, one participant who switched back their assessment to their initial belief stated, "*[I] disagree with the reasoning, but it convinced me that my initial assessment was correct*". In this case, the participant inferred that poor explanations may imply poor accuracy for AI assessments.

(a) **Difference between stage 2 and stage 5.** P-values from post-hoc Tukey HSD test for the *correct assessment, incorrect localization* scenario measuring localization accuracy from stage 2 compared to stage 5. There are statistically significant differences between strategies A to B and D for localization accuracy.

| ***Correct Assessment, Incorrect Localization*** Scenario<br>Metric: Localization Accuracy | |
|---|---|
| **Strategy Pairs** | **Tukey HSD p-value** |
| *A - B* | **0.019** |
| *A - D* | **0.019** |
| *A - F* | 0.275 |
| *B - D* | 0.900 |
| *B - F* | 0.607 |
| *D - F* | 0.607 |

(b) **Difference between stage 2 and stage 5.** P-values from post-hoc Tukey HSD test for the *incorrect assessment* scenario measuring assessment correctness from stage 2 compared to stage 5. There are statistically significant differences between strategies A to B, D, and F for the participant's assessment correctness.

| ***Incorrect Assessment*** Scenario<br>Metric: Assessment Accuracy | |
|---|---|
| **Strategy Pairs** | **Tukey HSD p-value** |
| *A - B* | **0.008** |
| *A - D* | **0.035** |
| *A - F* | **0.001** |
| *B - D* | 0.900 |
| *B - F* | 0.900 |
| *D - F* | 0.662 |

Table 3. Post-hoc Tukey HSD tests for determining statistically significant differences between the participant's localalization and assessment accuracy.

In contrast to strategy A, the accuracy for all other explanation strategies decreased as participants' switched their original correct assessments to agree with the AI's incorrect assessment in the *incorrect assessment* scenario. Strategy A shows a steep increase in accuracy from stage 3 to stage 4 (Figure 6), suggesting that the local text-based annotations revealed in stage 4 may have played a large role in helping participants calibrate their reliance on the AI.

*5.2.2 Incorrect localizations within causal explanations could lead to incorrect rationalizations.* In cases where participants were shown a correct assessment but an incorrect localization from the AI, causal explanations (strategy A) misled human localization more often than other explanation strategies. An ANOVA for the *correct assessment, incorrect localization* scenario indicated a significant difference in the impacts that different explanation strategies have on localization accuracy ($p < 0.01$).[4] A post-hoc Tukey HSD test revealed statistically significant differences in the Stage 5 - Stage 2 slopes between explanation strategies B and D versus A (see Table 3a). Strategy A shows a sharp drop in accuracy from stage 2 to 3 (Figure 6), suggesting that the presentation of local visual annotations, as part of a causal explanation, may have played a large role in encouraging inappropriate reliance upon incorrect AI localizations.

Interestingly, the presentation of the incorrect localization in stage 3 resulted in several participants changing the location of their pinpoint for strategy A, with a strong increase in overall confidence (Figure 6; third row, middle column). In this case, participants could have associated incorrect features (corresponding to the incorrect localization) with the correct damage assessment, potentially influencing how they made damage assessments later on in the study.

*5.2.3 Causal explanation strategies decrease over-reliance on AI.* We found that, when shown causal explanations (strategy A), humans relied less on the AI when the assessment was incorrect.

[4]See Table 8 in the Appendix B.3 for detailed results.

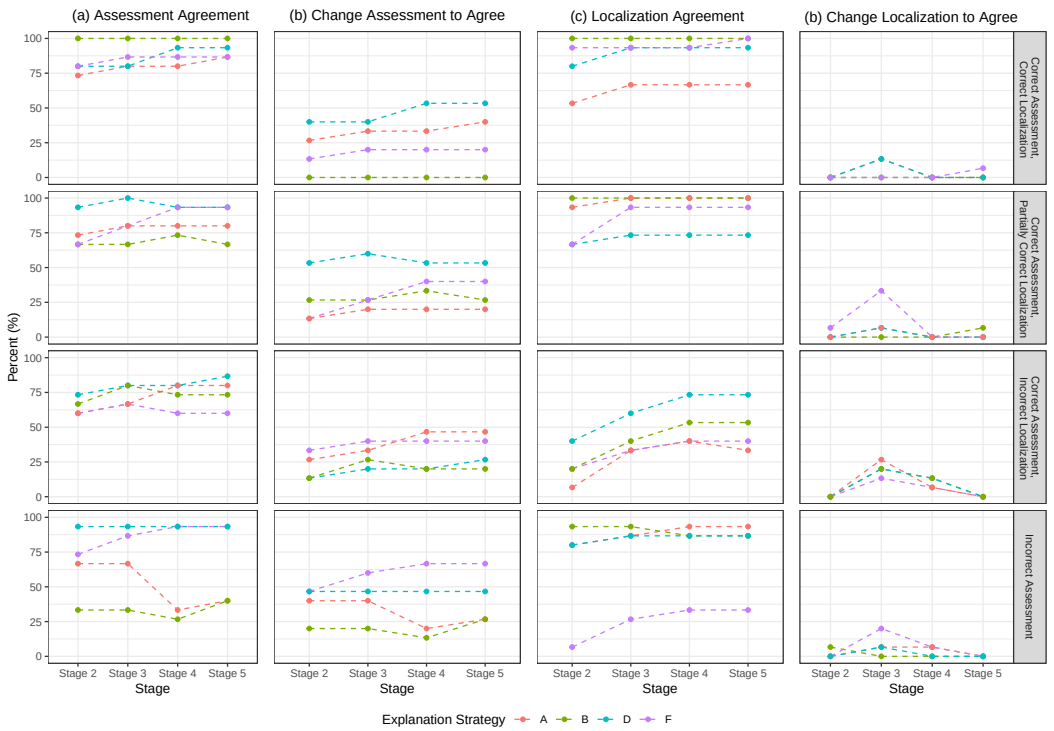



Fig. 7. Measurements to evaluate reliance on AI. The facet on the y-axis shows the four different scenarios that the participants were presented. The percent on the y-axis represents the ratio of participants (a) that agreed with the AI's assessment, (b) changed their assessment to agree with the AI's assessment, (c) that agreed with the AI's localization of the damage, and (d) that changed their pinpoint to agree with the AI's localization of the damage. The x-axis shows the five stages described in Figure 5

Drawing upon prior literature, we report multiple related behavioral measures to better understand participants' reliance on AI (**RQ3**) [59, 64]:

- *Assessment Agreement.* How often participants' assessments agree with the AI's assessment.
- *Change Assessment to Agree.* How often participants choose to change their assessment to agree with the AI's assessment.
- *Localization Agreement.* How often participants agree with the AI's localization of damage in the image (based on the location of the participant's pinpoint).
- *Change Localization to Agree.* How often participants choose to change the location of their pinpoint to agree with the AI's localization of damage.

An ANOVA for the *incorrect assessment* scenario showed a significant difference in the impact that different explanation strategies had on assessment agreement ($p < 0.05$)[5]. A post-hoc Tukey HSD test (Table 4) for this scenario revealed a statistically significant difference in the Stage 5 - Stage 2 slope between explanation strategy A (causal explanation) and B (comparing before and

[5]See Table 9 and Table 11 in Appendix B.3

Table 4. **Difference between stage 2 and stage 5.** P-values from post-hoc Tukey HSD test for the *incorrect assessment* scenario measuring assessment agreement from stage 2 compared to stage 5. There are statistically significant differences between strategies A to B and A to F for the participant's assessment agreement.

| ***Incorrect Assessment*** Scenario<br>Metric: Assessment Agreement | |
|---|---|
| **Strategy Pairs** | **Tukey HSD p-value** |
| *A - B* | **0.041** |
| *A - D* | 0.140 |
| *A - F* | **0.002** |
| *B - D* | 0.900 |
| *B - F* | 0.675 |
| *D - F* | 0.367 |

after features to explain a building damage assessment) as well as between explanation strategy A and F (identifying the number of damaged structures to explain a building damage assessment).

For strategy A in the *incorrect assessment* scenario where the AI localization is correct, we see a huge decline in assessment agreement when the text-based explanations are presented in stage 4. These trends are also observed in Figure 6. The increase in disagreement with the AI in stage 4 for strategy A could be a result participants judging that the causal arguments for the incorrect assessment actually do not make much sense. In contrast to strategies A and B, in the *incorrect assessment* scenario, strategy F shows an increase in reliance according to all four reported reliance metrics. For instance, before showing the AI's localization in stage 3, 6.7% of participants disagreed with the AI's localization. However, after seeing the explanations in stages 3 and 4, 33% of participants agreed with the AI's localization. Participants were misled by explanation strategy F to adopt the incorrect localization, resulting in a very low assessment accuracy as seen in Figure 6.

*5.2.4 Strategy type is less important when the AI is accurate and explanations are high quality.* We found no significant differences in assessment accuracy or reliance across explanation strategy types in the scenario when the AI assessment was correct. By contrast, as described above in Subsection 5.2.1, different explanation strategies had differential impacts on participants' assessment accuracy and reliance in scenarios where the AI was actually incorrect[6].

*5.2.5 Participants perceive explanations as most helpful when the AI is accurate.* Each stage in our study provided additional information from the AI to help the user make their damage assessment. Figure 5 shows the additional information that was presented to the users in each stage. In every stage except for Stage 1 (which does not present any information from the AI), we asked the user how helpful the additional information was in making their damage assessment. After assessing the damage, participants were asked to assess how helpful they found the additional information in each stage, using a 4-point Likert scale. Based on the additional information shown at a given stage (i.e., AI assessment, visual annotations, visual and text-based annotations, or global explanations), we asked participants questions such as, "How helpful was the AI assessment?" or "How helpful were the annotations?". Overall, our results suggest that participants generally found explanations most helpful in scenarios where the AI assessment and localization were correct. Our results also

[6]See Table 6 in Appendix B.3 for full results

suggest that participants generally found explanations less helpful when the AI's localization was incorrect or when the localization is correct and the assessment is incorrect. Figure 12 in Appendix B.3 describes more specific trends for this metric.

*5.2.6 Qualitative Analysis: Impact of AI Assessment and Explanations.* In this subsection, we report on participant responses to an optional question that was visible across every stage, asking participants whether and how they believed the AI affected their assessment. Since this question was optional, there was data from not every participant for every stage.

In general, participants stated that the AI either confirmed their initial damage assessment or, in the scenario where the localization was incorrect, that the AI did not highlight meaningful regions. Below we present examples of participant quotes, illustrating trends that emerged for each scenario.

Among participants who selected the correct damage assessment from the very beginning in the *correct assessment, correct localization* scenario, several noted that seeing the AI's assessment and explanations increased their confidence in their original decision. Participants who initially selected an incorrect damage assessment also described why they changed their assessment to match the AI's. For example, one participant who had originally selected the incorrect damage assessment changed to agree with the AI's assessment because "*...[the AI] came to the same conclusion and description of damage as me.*". A couple of participants reported that they chose to agree with the AI's assessment because the AI pointed out specific damage that they previously had not noticed. Interestingly, some participants noted that certain explanations were similar to their own line of reasoning. Four out of fifteen participants noted that the AI provides descriptions that match their own thought process for strategy B and two out of fifteen participants for strategy D.

By contrast, in the *correct assessment, partially correct localization* scenario, very few participants made comments about the AI reinforcing their assessment or providing similar reasoning. Instead, participants focused on pointing out pieces of information that the AI did not mention or realizing a damaged area/structure that they previously did not see. For example, one participant who originally had selected the incorrect assessment and later agreed with the AI said, "*It appears that I was looking at the wrong thing, the section that the AI has pointed out does have minor damage*". Another participant pointed out the flaws in the AI's assessment saying, "*[The AI] was correct about the damages. AI didn't notice the homes destroyed with what looks like only a foundation of a house after the disaster*".

In the *correct assessment, incorrect localization* scenario, participants were split between agreeing with or not agreeing with the AI's localization. Four out of fifteen participants for explanation strategy A, B, and D and three out of fifteen participants for explanation strategy F pointed out that the AI did not identify any meaningful regions. One participant said, "*[The] AI annotation is over an area that did not have damaged structure*". However, a few participants were misled by the incorrect annotations, resulting in behaviors such as moving the location of their pinpoint to agree with the AI's localization of damage or changing their damage assessment. One participant stated, "*[The AI] helped to direct the pin location instead of a general area*". Three participants for strategy F, three for strategy D, and two for strategy A made comments about the AI reinforcing their damage assessment and increasing their confidence, despite the AI being incorrect.

When the AI's damage assessment was actually *incorrect* and the participant's initial damage assessment matched the AI's assessment shown in Stage 2, several participants were incorrectly reassured by this and noted that the AI reinforced their original assessment. One participant said, "*It validated and reinforced my assessment. It gave me more confidence*". However, this participant changed their assessment to disagree with the AI and agree with the ground truth in stage 4 and stage 5 when shown detailed explanations from the AI about the annotations and overall assessment. The AI's incorrect assessment not only reassured participants who originally assessed

incorrectly; it also incorrectly influenced participants who originally assessed correctly. In one case when the AI predicted '*destroyed*' for an image that only presented '*minor damage*', the participant switched their correct assessment in Stage 1 to agree with the AI in stage 2 noting, "*The AI made me reconsider so I looked even harder for a destroyed building*". Interestingly, one participant that originally assessed the damage correctly and switched to agree with the AI's incorrect assessment then switched back to their original assessment because they disagreed with the reasoning of the AI's local explanations in Stage 4 stating, "*[I] disagree with the reasoning, but it convinced me that my initial assessment was correct*". The reasoning presented in the text-based annotations helped the participant understand the limitations of the AI. This example reflects a broader trend of strategy A (causal explanations) helping participants judge when *not* to agree with the AI in the *incorrect assessment* scenario, as shown in Figure 6.

## 6 DISCUSSION

Through a sequence of two online studies, we have elicited explanations from humans on a visual building damage assessment task and evaluated the impacts of different human explanation strategies on human accuracy and reliance upon AI-based assessments. Our findings offer insights into the kinds of explanation strategies humans employ when providing rationales for their decisions, in the context of visual decision-making tasks. We introduce a new approach for exploring the impacts that prospective explainable AI techniques might have on human-AI decision-making, by presenting participants with different types of human-generated explanations, framed as AI explanations. Using this approach, we investigated the impacts of different human explanation strategies on human accuracy and reliance upon AI-based assessments.

Overall, we found that compared with other explanation strategies, causal explanations misled humans less and served to decrease over-reliance on AI damage assessments. A wave of recent empirical results indicates that presenting AI explanations can often backfire—failing to improve or even *harming* human-AI decision-making in practice [4, 21, 25, 28, 30, 47]. For example, presenting explanations has sometimes been shown to promote over-reliance on AI recommendations, whether by lulling humans towards undeserved trust or by inducing cognitive overload [4, 30, 47]. In the context of this prior literature, our results help motivate the need to empirically investigate the impacts of different types of explanations on human-AI decision-making, across different real-world tasks and contexts. As others in the field have highlighted, AI explanations are not a monolith: there are many possible kinds of explanations, which may have different (potentially context-dependent) effects on human-AI decision-making [37, 41, 61].

In our study, we found that causal explanations had unique impacts among the explanation strategies we investigated: **in the context of erroneous AI damage assessments, causal explanations empowered humans to correctly second-guess the AI.** Our findings suggest that humans are better at calibrating their reliance on AI assessment by reasoning about causal arguments, versus by assessing other kinds of explanations. This interpretation aligns with prior research on human causal cognition, which suggests that humans are predisposed to reason about the world in a causal, rather than purely statistical or associative manner. Indeed, some errors and fallacies that have been observed in human probabilistic or statistical reasoning can be understood as symptoms of this tendency: instances where humans attempt to use causal reasoning and assumptions in situations where these do not apply [3, 20, 29]. For these reasons, it may be that humans are better able to spot faulty reasoning in causal explanations versus other explanation strategies.

Although causal explanations helped humans identify erroneous AI assessments of the *extent* of damage, we found that **causal explanations led humans astray when the AI incorrectly identified the location of the damage**. Interestingly, our findings suggest that participants may

have updated their own line of reasoning after seeing only the *visual annotation* components of causal explanations, presented in Stage 3. The presentation of visual annotations did not have this effect for other explanation strategies, suggesting that the types of visual annotations that human explainers generate in the context of causal explanations are in themselves more persuasive than those associated with other strategies. It may be, for example, that the visual annotations associated with human-generated causal explanations tend to visually highlight potential causal factors, which influence human damage localization.

Interestingly, we found that in our context of building damage assessment from satellite imagery, causal explanations were most helpful in improving decision-making in cases where the AI's damage assessment was incorrect. By contrast, **when the AI's assessment was correct, the type of explanation strategy presented did not significantly impact participants' accuracy or their reliance on AI assessments.** This suggests that causal explanations may be particularly valuable in settings where AI systems are likely to be highly imperfect, including high uncertainty settings or cases where an AI system is particularly vulnerable to blindspots [7, 27, 31].

## 6.1 Limitations

In this section, we briefly highlight key limitations across both of our studies:

*6.1.1 "Who is the explanation for?"* In Study 1, participants were asked to generate annotations and rationalizations for their damage assessment with the goal of convincing another person that their answer is correct. However, recent discussions in the human-centered explainable AI literature emphasize the importance of tailoring explanations to particular stakeholder groups (i.e., model developers, business owners, frontline decision-makers, decision subjects, regulatory bodies), who may have different use cases for AI explanations or different expertise through which to interpret explanations (e.g., [25, 35]). In our study, we asked participants to convince "another person." However, it is possible that presenting a more specific prompt in Study 1 (e.g., explicitly specifying that the user of the explanation would be another participant on the Prolific platform) would have yielded a different distribution of explanation strategies.

*6.1.2 Potential Effects of Image Selection on the Identified Human Explanation Strategies.* We filtered the xBD dataset to only use image pairs with one damaged structure or multiple damaged structures with the same damage assessment. This constraint limited the number of image we had represented in the dataset for each natural disaster. This also could have impacted the distribution of explanation strategies that we observed in our study. For instance, it is unclear how and to what extent the explanation strategies might differ when participants are presented with images including several damaged regions with different damage assessments.

*6.1.3 Instructed to annotate, not localize damage.* The participants in Study 1 were not explicitly instructed to annotate the specific building they thought was damaged. They were only advised to mark evidence relevant to the level of damage that they wanted to argue for. Therefore, it is possible that some Study 1 participants could have annotated evidence of damage to surrounding areas, without annotating the specific building that they believed was damaged. When selecting images for the *correct assessment, correct localization* scenario in Study 2, we only considered observations in which a damaged building was clearly marked. This limited the number of observations we could choose from for the *correct assessment, correct localization* scenario.

*6.1.4 Availability of Explanations from Study 1.* We wanted to evaluate different explanation strategies and different variations of explanation strategies. However, not all explanation strategies

were represented equally for each image set[7]. In addition, in order to evaluate one explanation strategy at a time in Study 2, we were limited to observations from Study 1 that only used one explanation strategy. Due to these constraints, most of the image pairs within each scenario are different for each explanation strategy. This meant that it was not feasible to assess the impacts of particular explanation strategies independently of particular image pairs. Thus, to control for baseline differences across image pairs (i.e., differences at Stage 1, before any AI outputs were shown to participants), our analyses compared *changes* in particular metrics (e.g., accuracy and confidence) across stages, rather than comparing absolute values. Nonetheless, it is possible that we were still unable to observe certain effects due to differences across image pairs. For example, if participants' baseline accuracy for a given image pair was near 100% for a given scenario[8], then this may have masked explanation strategy effects that we would have otherwise observed (i.e., due to a ceiling effect).

Some participants noted in the end-of-task survey that some of the images were low quality or hard to decipher due to the satellite images being taken at different times of the day or different seasons. Unfortunately, the quality of the images is typical in this domain and also a limitation of the open-source data set [18].

## 7 FUTURE WORK

There are many opportunities for further work to build upon our bottom-up taxonomy of human explanation strategies. We plan to prepare and release an open-source dataset mapping explanation strategies to specific examples of human explanations that we collected in our study. This dataset will allow researchers to explore ways to generate explanations that associate with a certain explanation strategy for image classification tasks. Generating these explanation strategies will allow researchers to evaluate to what extent the explanation strategies impact a decision makers' accuracy and reliance on AI at a larger scale.

Future work should consider running Study 2 at a larger scale. As discussed in our Limitations, in this study we were only able to evaluate one explanation strategy per image set. Future studies should consider evaluating all four explanation strategies on the same image, to better separate out potential impacts of particular image sets versus explanation strategies. Furthermore, while this study focused on crowdworkers, researchers should consider recruiting subject matter experts to see how different explanation strategies impact their workflow and whether they have a preference for a particular strategy. Overall, XAI for building damage assessment and disaster relief remains a critical yet underexplored research area. Future work is greatly needed to better understand how AI-based decision supports can be designed effectively for this context.

Since the core explanation strategies we identified can be generalized to other visual decision-making domains, such as radiology, we encourage researchers to explore the impact that visual annotations paired with text-based annotations have on human accuracy and reliance on AI in a broader range of contexts. For example, in medical imaging, the majority of current explainable AI techniques focus on providing saliency maps [48]. The kinds of human-generated visual and text-based explanations presented to participants in our study are much richer by comparison.

## 8 CONCLUSION

Explainable AI techniques are increasingly being evaluated to understand how they aid practitioners during the decision making process. Findings from numerous user studies call for a more human-centered approach to explainable AI for human-AI decision making. To address the need for

[7] See table 9 in Appendix A.2 to see the distribution of explanation strategies from Study 1.

[8] See assessment accuracy for *correct assessment, correct localization* scenario in Section 5.2, Figure 6

human-centered explainable AI and understand its impact on the decision making process, we conducted a series of two studies to identify explanation strategies that humans generate during visual decision making tasks and to understand how presenting these explanation strategies to human decision-makers impacts their accuracy and reliance upon AI.

We identified four core explanation strategies in the context of building damage assessment from satellite imagery: causal arguments, before-and-after comparison, the proportion of structure damage, and the number of structures damaged. Based on those four core explanations, we evaluate whether and how they impact humans' assessment accuracy and reliance on AI assessments. Our results show that causal explanations can help humans appropriately calibrate their reliance on AI damage assessments in cases where the AI is incorrect. However, causal explanations can also lead humans astray when the AI localization is incorrect. As causal explanation strategies are applicable across a broad range of real-world domains beyond damage assessment or other visual decision-making tasks, our results suggest new guidance on how to make explanations more useful and effective in practice.

## REFERENCES


[1] Amina Adadi and Mohammed Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

[2] Josh Andres, Christine T. Wolf, Sergio Cabrero Barros, Erick Oduor, Rahul Nair, Alexander Kjærum, Anders Bech Tharsgaard, and Bo Schwartz Madsen. 2020. Scenario-based XAI for Humanitarian Aid Forecasting. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, 1–8. https://doi.org/10.1145/3334480.3382903

[3] Joseph L Austerweil and Thomas L Griffiths. 2011. Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science* 35, 3 (2011), 499–526.

[4] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. *Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance.* Association for Computing Machinery, 1–16. https://doi.org/10.1145/3411764.3445717

[5] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. (2012).

[6] Andrea Brennen. 2020. What Do People Really Want When They Say They Want "Explainable AI?" We Asked 60 Stakeholders. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, 1–7. https://doi.org/10.1145/3334480.3383047

[7] Ángel Alexander Cabrera, Abraham J Druck, Jason I Hong, and Adam Perer. 2021. Discovering and Validating AI Errors With Crowdsourced Failure Reports. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–22.

[8] 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 (Nov 2019), 1–24. https://doi.org/10.1145/3359206

[9] David S Channin, Pattanasak Mongkolwat, Vladimir Kleper, and Daniel L Rubin. 2009. The annotation and image mark-up project. https://doi.org/10.1148/radiol.2533090135

[10] Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are Visual Explanations Useful? A Case Study in Model-in-the-Loop Prediction. (Jul 2020). https://arxiv.org/abs/2007.12248v1 arXiv: 2007.12248v1.

[11] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A Case for Humans-in-the-Loop: Decisions in the Presence of Erroneous Algorithmic Scores. (Feb 2020). https://doi.org/10.1145/3313831.3376638

[12] Abhirup Dikshit and Biswajeet Pradhan. 2021. Interpretable and explainable AI (XAI) model for spatial drought prediction. *Science of The Total Environment* 801 (Dec 2021), 149797. https://doi.org/10.1016/j.scitotenv.2021.149797

[13] Upol Ehsan, Brent Harrison, Larry Chan, and Mark O. Riedl. 2018. Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18)*. ACM, 81–87. https://doi.org/10.1145/3278721.3278736

[14] Upol Ehsan and Mark O. Riedl. 2020. Human-Centered Explainable AI: Towards a Reflective Sociotechnical Approach. In *HCI International 2020 - Late Breaking Papers: Multimodality and Intelligence (Lecture Notes in Computer Science)*, Constantine Stephanidis, Masaaki Kurosu, Helmut Degen, and Lauren Reinerman-Jones (Eds.). Springer International Publishing, 449–466. https://doi.org/10.1007/978-3-030-60117-1_33

[15] Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O. Riedl. 2021. *Operationalizing Human-Centered Perspectives in Explainable AI.* Association for Computing

Machinery, 1–6. https://doi.org/10.1145/3411763.3441342

[16] Ziba Gandomkar and Claudia Mello-Thoms. 2019. Visual search in breast imaging. *The British journal of radiology* 92, 1102 (2019), 20190057.

[17] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual Visual Explanations. *arXiv:1904.07451 [cs, stat]* (Jun 2019). http://arxiv.org/abs/1904.07451 arXiv: 1904.07451.

[18] Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. 2019. Creating xBD: A dataset for assessing building damage from satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 10–17.

[19] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating Visual Explanations. In *Computer Vision – ECCV 2016 (Lecture Notes in Computer Science)*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, 3–19. https://doi.org/10.1007/978-3-319-46493-0_1

[20] Ralph Hertwig and Gerd Gigerenzer. 1999. The 'conjunction fallacy' revisited: How intelligent inferences look like reasoning errors. *Journal of behavioral decision making* 12, 4 (1999), 275–305.

[21] Kenneth Holstein and Vincent Aleven. 2021. Designing for human-AI complementarity in K-12 education. *arXiv preprint arXiv:2104.01266* (2021).

[22] Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 375–385.

[23] Neal Jean, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. Combining satellite imagery and machine learning to predict poverty. *Science* 353, 6301 (Aug 2016), 790–794. https://doi.org/10.1126/science.aaf7894

[24] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. 2019. XRAI: Better Attributions Through Regions. https://arxiv.org/abs/1906.02825 arXiv: 1906.02825.

[25] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Jennifer Wortman Vaughan. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[26] Asim B. Khajwal and Arash Noshadravan. 2021. An uncertainty-aware framework for reliable disaster damage assessment via crowdsourcing. *International Journal of Disaster Risk Reduction* 55 (Mar 2021), 102110. https://doi.org/10.1016/j.ijdrr.2021.102110

[27] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[28] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. *arXiv preprint arXiv:2112.11471* (2021).

[29] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2017. Building machines that learn and think like people. *Behavioral and brain sciences* 40 (2017).

[30] Himabindu Lakkaraju and Osbert Bastani. 2020. " How do I fool you?" Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 79–85.

[31] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. 2017. Identifying unknown unknowns in the open world: Representations and policies for guided exploration. In *Thirty-first aaai conference on artificial intelligence*.

[32] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T. Freeman, Phillip Isola, Amir Globerson, Michal Irani, and Inbar Mosseri. 2021. Explaining in Style: Training a GAN to explain a classifier in StyleSpace. *arXiv preprint arXiv:2104.13369* (2021).

[33] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[34] 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 (CHI '20)*. Association for Computing Machinery, 1–15. https://doi.org/10.1145/3313831.3376590

[35] Q. Vera Liao and Kush R. Varshney. 2021. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. (Dec 2021). http://arxiv.org/abs/2110.10790 arXiv: 2110.10790.

[36] Zhong Qiu Lin, Mohammad Javad Shafiee, Stanislav Bochkarev, Michael St Jules, Xiao Yu Wang, and Alexander Wong. 2019. Do explanations reflect decisions? A machine-centric strategy to quantify the performance of explainability algorithms. *arXiv preprint arXiv:1910.07387* (2019).

[37] Zachary C Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[38] Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. (Nov 2017). http://arxiv.org/abs/1705.07874 arXiv: 1705.07874.

[39] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 4768–4777. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[40] Sahar S. Matin and Biswajeet Pradhan. 2021. Earthquake-Induced Building-Damage Mapping Using Explainable AI (XAI). *Sensors* 21, 13 (2021). https://doi.org/10.3390/s21134489

[41] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[42] Christoph Molnar. 2019. *Model-Agnostic Methods*. Lulu.

[43] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D. Ragan. 2019. The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems. 7 (Oct 2019), 97–105. https://ojs.aaai.org/index.php/HCOMP/article/view/5284

[44] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. 2017. Feature Visualization. *Distill* 2, 11 (Nov 2017), e7. https://doi.org/10.23915/distill.00007

[45] Barak Oshri, Annie Hu, Peter Adelson, Xiao Chen, Pascaline Dupas, Jeremy Weinstein, Marshall Burke, David Lobell, and Stefano Ermon. 2018. Infrastructure Quality Assessment in Africa using Satellite Imagery and Deep Learning. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Jul 2018), 616–625. https://doi.org/10.1145/3219819.3219924

[46] Bhavik N Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, et al. 2019. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ digital medicine* 2, 1 (2019), 1–10.

[47] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–52.

[48] Mauricio Reyes, Raphael Meier, Sérgio Pereira, Carlos A Silva, Fried-Michael Dahlweid, Hendrik von Tengg-Kobligk, Ronald M Summers, and Roland Wiest. 2020. On the interpretability of artificial intelligence in radiology: challenges and opportunities. *Radiology: Artificial Intelligence* 2, 3 (2020), e190043.

[49] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. 1135–1144. https://doi.org/10.1145/2939672.2939778

[50] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. (Aug 2016). http://arxiv.org/abs/1602.04938 arXiv: 1602.04938.

[51] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G. Blankenberg, Andrew Y. Ng, Matthew P. Lungren, and Pranav Rajpurkar. 2021. Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. *medRxiv* (Mar 2021). https://doi.org/10.1101/2021.02.28.21252634v1

[52] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2020. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision* 128, 2 (Feb 2020), 336–359. https://doi.org/10.1007/s11263-019-01228-7

[53] Rita Sevastjanova, Fabian Beck, Basil Ell, Cagatay Turkay, Rafael Henkin, Miriam Butt, Daniel A Keim, and Mennatallah El-Assady. 2018. Going beyond visualization: Verbalization as complementary medium to explain machine learning models. (Oct 2018).

[54] Hidehiko Shishido, Koyo Kobayashi, Yoshinari Kameda, and Itaru Kitahara. 2021. Method to Generate Building Damage Maps by Combining Aerial Image Processing and Crowdsourcing. *Journal of Disaster Research* 16, 5 (Aug 2021), 827–839. https://doi.org/10.20965/jdr.2021.p0827

[55] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. (Apr 2014). http://arxiv.org/abs/1312.6034 arXiv: 1312.6034.

[56] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 180–186.

[57] Richard Szeliski. 2010. *Computer vision: algorithms and applications*. Springer Science & Business Media.

[58] Defense Innovation Unit. 2021. U.S. Government and Nonprofit Organization Host Prize Competition to Leverage the Latest Technology to Detect and Defeat Illegal Fishing. https://www.diu.mil/latest/us-government-and-nonprofit-organization-host-prize-competition-xview3

[59] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (Oct 2021), 327:1–327:39. https://doi.org/10.1145/3476068

[60] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2018. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *arXiv:1711.00399 [cs]* (Mar 2018). http://arxiv.org/abs/1711.00399 arXiv: 1711.00399.

[61] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[62] Xinru Wang and Ming Yin. 2021. *Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making*. Association for Computing Machinery, 318–328. https://doi.org/10.1145/3397481.3450650

[63] Christopher Yeh, Anthony Perez, Anne Driscoll, George Azzari, Zhongyi Tang, David Lobell, Stefano Ermon, and Marshall Burke. 2020. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications* 11, 1 (May 2020), 2583. https://doi.org/10.1038/s41467-020-16185-w

[64] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. (Jan 2020). https://doi.org/10.1145/3351095.3372852

[65] Alexandra Zytek, Dongyu Liu, Rhema Vaithianathan, and Kalyan Veeramachaneni. 2021. Sibyl: Understanding and Addressing the Usability Challenges of Machine Learning In High-Stakes Decision Making. (Sep 2021). http://arxiv.org/abs/2103.02071 arXiv: 2103.02071.

# A APPENDIX - STUDY 1

## A.1 Study 1 Image Set

The image sets used in Study 1 is provided on our temporarily anonymous GitHub page[9]. The training data used in both Study 1 and Study 2 can found here[10] and the main data used in Study 1 can be found here[11].

## A.2 Study 1 Detailed Analyses

The accuracy for each image set is presented in Figure 8. Some of the images were harder for the participants to determine the correct assessment than others. The two image sets with the highest accuracy are wildfires where the building in the image was completely burnt to the ground. The two hardest images, Midwest Flooding 429 and Mexico Earthquake 8, for participants to assess contained multiple structures in the image making it slightly more difficult for them to find the one structure that was damaged. It is also notable that none of the participants had correct localization for those two images.

Two coders reviewed the results from study 1 and assigned strategies to every response. Some responses had multiple strategies assigned. The differences between the number of strategies the two coders assigned for each sub-strategy are shown in Figure 9. As seen in Figure 10, when the participant got the assessment correct, the most prevalent code across all image sets is was code B. The other top prevalent codes include codes A, D, and F.

The percent of observations per image set where participants cited their local explanations within their global explanations seen in Figure 11.

[9]https://tmp-cscw2022.github.io/
[10]https://tmp-cscw2022.github.io/TrainingData.html
[11]https://tmp-cscw2022.github.io/MainData/MainData.html

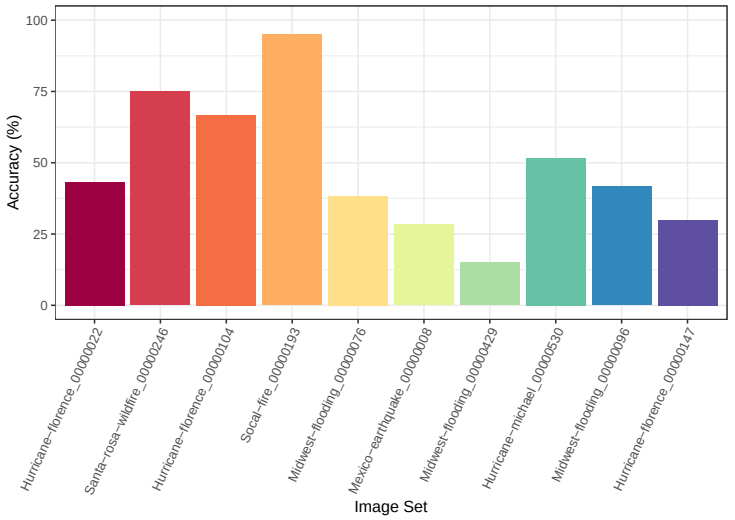



Fig. 8. Accuracy for each image set from study 1. Accuracy is calculated by the number of participants who got the damage assessment correct out of all participant. The images were shown in a random order to minimize ordering effects.

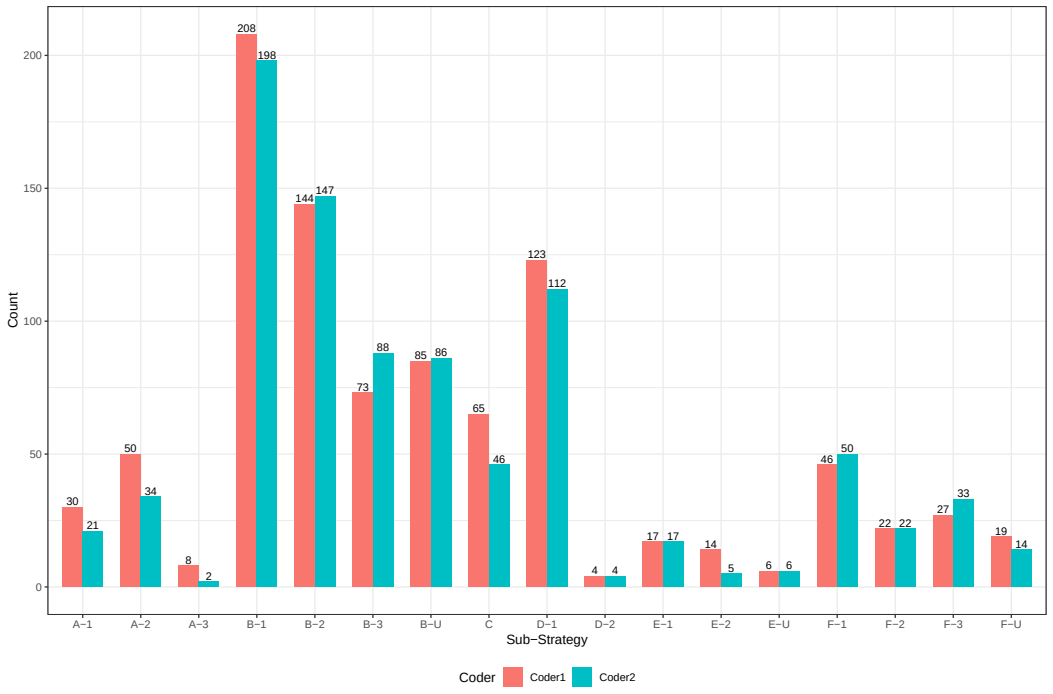



Fig. 9. Frequency of the strategies assigned by each coder. For both coders, strategy B appeared the most, followed by D, F, A, C, and E

## B APPENDIX - STUDY 2

Detailed documentation for this study can be viewed on our temporarily anonymous GitHub site here[12]. Our interface codebase is available to those who wish to replicate our study (link hidden for anonymity). Sections below point to more specific pages on the GitHub site for easier navigation as well as interpretations of more detailed analyses.

### B.1 Training Phase 2 Data

Details for how the data for the second training phase can be found on our temporarily, anonymous GitHub site here[13]. This data was solely used to help participants get familiarized with the task and the new information they are provided in each stage.

### B.2 Representative Explanations

The method for the data set used in the second study is detailed in Section 5.1.2. More detailed documentation on the selection method can be found here[14] and the final data set used for the study can be seen here[15].

[12]https://tmp-cscw2022.github.io/
[13]https://tmp-cscw2022.github.io/WalkthroughData.html
[14]https://tmp-cscw2022.github.io/representativedataset.html
[15]https://tmp-cscw2022.github.io/RepresentativeExplanations/Representative%20Explanations%20a30337b3cc864808882c0898ba4a537c.html

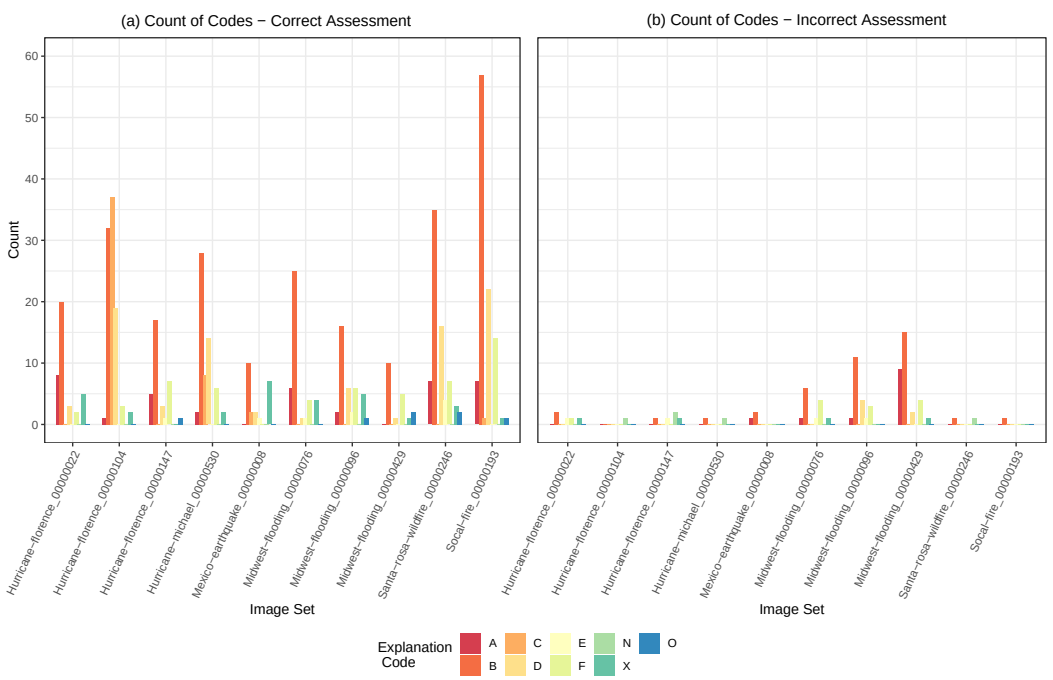



Fig. 10. Number of codes identified for each image set for participants that selected the correct damage assessment (a) and for each image set for participants that selected the incorrect damage assessment (b). Some participants included multiple codes for an image set.

Table 5. P-values from ANOVA single-factor test for each scenario for participant's assessment accuracy between stage 1 and stage 5.

| Metric: Assessment Accuracy | |
|---|---|
| **Scenario** | **ANOVA p-value** |
| *Correct Assessment, Correct Localization* | **0.005** |
| *Correct Assessment, Partially Correct Localization* | 0.24 |
| *Correct Assessment, Incorrect Localization* | 0.46 |
| *Incorrect Assessment* | 0.67 |

## B.3 Study 2 Detailed Analyses

Below we provide all of the p-values from our ANOVA tests.

We calculated similar ANOVA tests to determine if the change in the participants' localization accuracy between stage 1 and stage 5 was statistically significant, however, no scenario had a *p-value* less than 0.05 (Table 7 in Appendix Section B.3). This is to be expected as the participants' localization accuracy throughout the stages in the *correct assessment, incorrect localization* scenario had very small increases/decreases.

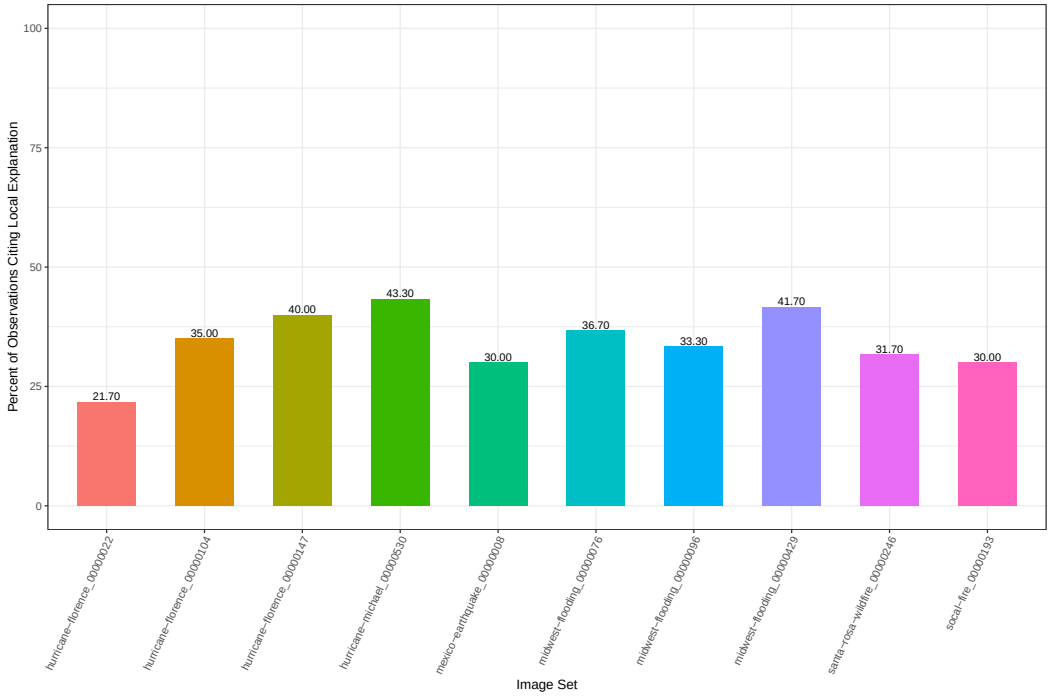



Fig. 11. Percent of observations per image set where participants cited local explanations.

Table 6. P-values from ANOVA single-factor test for each scenario for participant's assessment accuracy between stage 2 and stage 5.

| Metric: Assessment Accuracy | |
|---|---|
| **Scenario** | **ANOVA p-value** |
| *Correct Assessment, Correct Localization* | 0.511 |
| *Correct Assessment, Partially Correct Localization* | 0.089 |
| *Correct Assessment, Incorrect Localization* | 0.456 |
| *Incorrect Assessment* | **0.001** |

Table 7. P-values from ANOVA single-factor test for each scenario for participants localization accuracy between stage 5 and stage 1.

| Metric: Localization Accuracy | |
|---|---|
| **Scenario** | **ANOVA p-value** |
| *Correct Assessment, Correct Localization* | 0.30 |
| *Correct Assessment, Partially Correct Localization* | 0.40 |
| *Correct Assessment, Incorrect Localization* | 0.50 |
| *Incorrect Assessment* | 0.31 |

Table 8. P-values from ANOVA single-factor test for each scenario for participants localization correctness between stage 5 and stage 2.

| Metric: Localization Accuracy | |
|---|---|
| **Scenario** | **ANOVA p-value** |
| *Correct Assessment, Correct Localization* | 0.289 |
| *Correct Assessment, Partially Correct Localization* | 0.400 |
| *Correct Assessment, Incorrect Localization* | **0.01** |
| *Incorrect Assessment* | 0.273 |

Table 9. P-values from ANOVA single-factor test for each scenario for participant's assessment agreement between stage 1 and stage 5.

| Metric: Assessment Agreement | |
|---|---|
| **Scenario** | **ANOVA p-value** |
| *Correct Assessment, Correct Localization* | **0.005** |
| *Correct Assessment, Partially Correct Localization* | 0.24 |
| *Correct Assessment, Incorrect Localization* | 0.46 |
| *Incorrect Assessment* | **0.026** |

Table 10. **Difference between stage 1 and stage 5.** P-values from post-hoc Tukey HSD test for the *incorrect assessment* scenario measuring assessment agreement change from stage 1 compared to stage 5. There is a statistically significant difference between strategies A and F for the participant's assessment correctness.

| ***Incorrect Assessment*** Scenario<br>Metric: Assessment Agreement | |
|---|---|
| **Strategy Pairs** | **Tukey HSD p-value** |
| *A - B* | 0.871 |
| *A - D* | 0.264 |
| *A - F* | **0.023** |
| *B - D* | 0.667 |
| *B - F* | 0.132 |
| *D - F* | 0.667 |

*B.3.1 Humans did not significantly change their localizations to agree with the AI.* Whether the AI's localization was correct or incorrect, we do not find any significant difference in the impacts that different explanation strategies have on localization agreement. However, an ANOVA for the *correct assessment, partially correct localization* scenario ($p < 0.05$) indicated a difference in the impacts

Table 11. P-values from ANOVA single-factor test for each scenario for participant's assessment agreement between stage 2 and stage 5.

| Metric: Assessment Agreement | |
|---|---|
| **Scenario** | **ANOVA p-value** |
| *Correct Assessment, Correct Localization* | **0.005** |
| *Correct Assessment, Partially Correct Localization* | 0.24 |
| *Correct Assessment, Incorrect Localization* | 0.46 |
| *Incorrect Assessment* | **0.003** |

Table 12. P-values from ANOVA single-factor test for each scenario for participants localization agreement between stage 5 and stage 1.

| Metric: Localization Agreement | |
|---|---|
| **Scenario** | **ANOVA p-value** |
| *Correct Assessment, Correct Localization* | 0.511 |
| *Correct Assessment, Partially Correct Localization* | **0.020** |
| *Correct Assessment, Incorrect Localization* | 0.836 |
| *Incorrect Assessment* | 0.644 |

Table 13. **Difference between stage 1 and stage 5**. P-values from post-hoc Tukey HSD test for the *correct assessment, partially correct localization* scenario measuring localization agreement from stage 1 compared stage 5. There is a statistically significant difference between strategies B and F for localization agreement.

| ***Correct Assessment, Partially Correct Localization*** Scenario<br>Metric: Localization Agreement | |
|---|---|
| **Strategy Pairs** | **Tukey HSD p-value** |
| *A - B* | 0.90 |
| *A - D* | 0.90 |
| *A - F* | 0.09 |
| *B - D* | 0.90 |
| *B - F* | **0.02** |
| *D - F* | 0.09 |

that different explanation strategies have on localization agreement[16]. A post-hoc Tukey HSD test revealed statistically significant differences in the Stage 5 - Stage 1 slope between explanation strategy B and F.

[16]See Table 12 in Appendix B.3 for full results.

Table 14. P-values from ANOVA single-factor test for each scenario for participants localization agreement between stage 5 and stage 2.

| Metric: Localization Agreement | |
|---|---|
| **Scenario** | **ANOVA p-value** |
| *Correct Assessment, Correct Localization* | 0.511 |
| *Correct Assessment, Partially Correct Localization* | 0.084 |
| *Correct Assessment, Incorrect Localization* | 0.836 |
| *Incorrect Assessment* | 0.071 |

Explanation strategy F shows a positive trend from stage 2 to stage 4 for the *assessment agreement* and *change assessment to agree* metrics in the *correct assessment, partially correct localization* scenario. 33% of participants who saw the AI's localization in stage 3 for explanation strategy F changed their localization to agree with the AI's localization. However, the localization accuracy for strategy F in Figure 6 remained approximately 25% for all of the stages showing the partially incorrect localization did not help the participant identify the localization correctly.

Figure 12 visualizes the responses to the helpfulness question (on 4-point Likert scale from very unhelpful to very helpful) for every code, stage, and scenario.

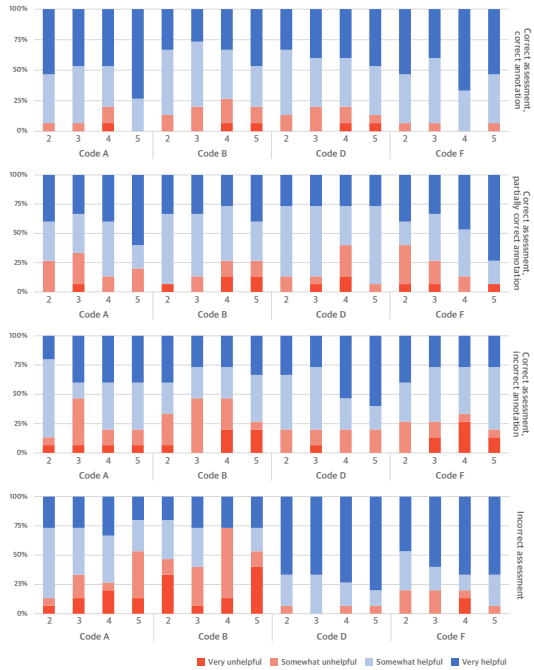



Fig. 12. Visualizing how helpful explanations in each stage and scenario were to participants for every code.