Eye into AI: Evaluating the Interpretability of Explainable AI Techniques through a Game With a Purpose

KATELYN MORRISON, Carnegie Mellon University, USA
MAYANK JAIN, Carnegie Mellon University, USA
JESSICA HAMMER, Carnegie Mellon University, USA
ADAM PERER, Carnegie Mellon University, USA

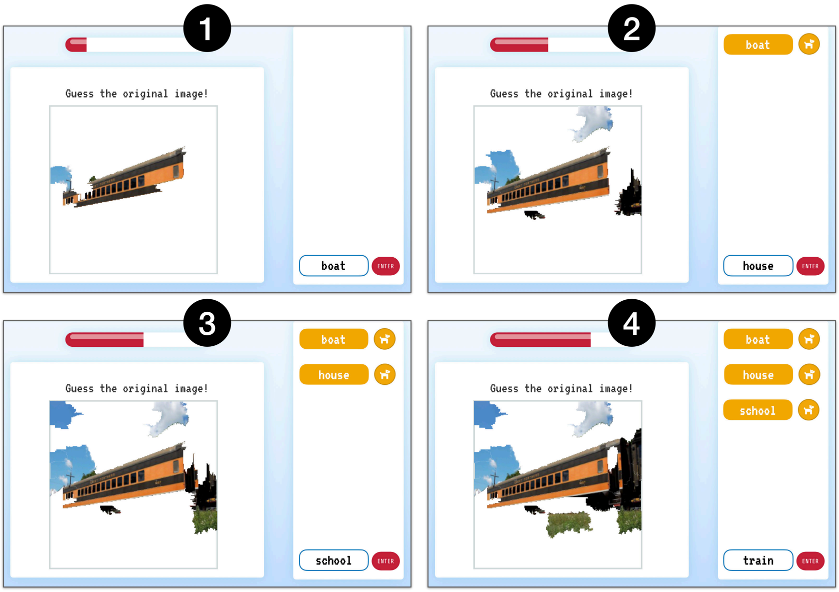

1 Guess the original image! boat
2 Guess the original image! house
3 Guess the original image! school
4 Guess the original image! train

Fig. 1. An example scenario of the guesser round in Eye into AI, where the player is shown an AI explanation of an image classified as a train. In 1), the player is shown the most salient portions of the image according to LIME (which we call the top 1 explanation for LIME), and the player incorrectly guesses a ‘boat‘. In 2), the next most salient portion of the image is revealed (top 2 explanation for LIME), leading the player to guess a ‘house‘. In 3), the top 3 explanation for LIME is shown, leading to an incorrect guess of ‘school‘. Finally, in 4), the top 4 explanation for LIME is revealed, leading the player to correctly guess ‘train‘.

Recent developments in explainable AI (XAI) aim to improve the transparency of black-box models. However, empirically evaluating the interpretability of these XAI techniques is still an open challenge. The most common evaluation method is algorithmic performance, but such an approach may not accurately represent

Authors’ addresses: Katelyn Morrison, kcmorris@cs.cmu.edu, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; Mayank Jain, mayankj@andrew.cmu.edu, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; Jessica Hammer, hammerj@andrew.cmu.edu, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA; Adam Perer, adamperer@cmu.edu, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.

2573-0142/2023/-ART
https://doi.org/

how interpretable these techniques are to people. A less common but growing evaluation strategy is to leverage crowd-workers to provide feedback on multiple XAI techniques to compare them. However, these tasks often feel like work and may limit participation. We propose a novel, playful, human-centered method for evaluating XAI techniques: a Game With a Purpose (GWAP), Eye into AI, that allows researchers to collect human evaluations of XAI at scale. We provide an empirical study demonstrating how our GWAP supports evaluating and comparing the agreement between three popular XAI techniques (LIME, Grad-CAM, and Feature Visualization) and humans, as well as evaluating and comparing the interpretability of those three XAI techniques applied to a deep learning model for image classification. The data collected from Eye into AI offers convincing evidence that GWAPs can be used to evaluate and compare XAI techniques.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**; • **Computing methodologies** → *Artificial intelligence.*

Additional Key Words and Phrases: Games With a Purpose, Explainable AI, Interpretability

ACM Reference Format:
Katelyn Morrison, Mayank Jain, Jessica Hammer, and Adam Perer. 2023. Eye into AI: Evaluating the Interpretability of Explainable AI Techniques through a Game With a Purpose. *Proc. ACM Hum.-Comput. Interact.* , CSCW (2023), 22 pages.

1 INTRODUCTION

Explainable AI (XAI) offers the promise of transparency and the ability to simplify complex AI models to be interpretable to humans. Transparency and simplicity matter in every phase of the development of AI systems. AI developers need techniques to ensure that their models align with their goals; stakeholders need oversight to ensure that the models they deploy behave as expected; and consumers, the end users of such models, need tools to help them appropriately trust AI when making decisions. However, while many XAI techniques have been proposed, the few that have been evaluated beyond anecdotal evidence suggest that end-users over-rely or under-rely on AI [6, 35, 61]. The field needs techniques to accurately assess how interpretable AI explanations are to humans and to understand where humans and AI explanations agree, using verifiable measurements that are deployed at scale with actual people. We argue that explanations cannot be evaluated in an automated manner, as only people can determine if an explanation makes sense for people.

In this paper, we present Eye Into AI, a scalable platform to support valid and verifiable assessments of AI explanations with humans: integrating XAI with games with a purpose (GWAPs). GWAPs are games designed for humans to play online that generate usable data as a by-product of gameplay. The GWAP method has been shown to be highly effective in collecting data from large populations online and validating that data at scale [57]. GWAPs have several advantages over conventional crowd image labeling, such as not relying on financial incentives [10], interesting audiences who do not enjoy crowd work but enjoy games [55], and improving engagement with higher persistence [26].

We argue that explainability is an appropriate and a challenging topic to tackle with GWAP methods. XAI is an appropriate fit for GWAP efforts because it requires human evaluation at scale for a task that can be broken down into pieces. Most GWAPs rely on collecting large amounts of data and then using validation to exclude low-quality data. However, common validation strategies are inappropriate for explainability tasks [17]. For example, in agreement designs, players' contributions are evaluated based on how similar they are to other players' inputs on the same task. For explainability, however, thoughtful and creative interpretations are needed to accurately assess if they are sensible. Therefore, we instead take an intrinsic validation approach, where the gameplay choices are closely aligned with the provision of high-quality human subject data [17]. For our problem, this means that providing thoughtful responses to ambiguous visual stimuli is the optimal way to play.

Specifically, we developed Eye into AI as an evaluation tool that researchers can use to compare and improve XAI techniques. To determine whether Eye into AI can be a successful evaluation tool for XAI, we address several open research questions:

- **RQ1.** How can Eye into AI help researchers understand the agreement between XAI techniques and humans? As a result, Eye into AI can provide data to help researchers improve and optimize their XAI techniques.
- **RQ2.** How can Eye into AI help researchers identify which technique is more interpretable than another? As a result, researchers will be able to assess which XAI techniques lead to better decisions.
- **RQ3.** How scalable can Eye into AI be for evaluating the interpretability of XAI techniques and the agreement between interpretability techniques and humans? The answer to t his will confirm the scalability of this technique for researchers interested in evaluating and comparing XAI techniques using human subjects.

To address these research questions, we conducted an initial empirical study on Eye into AI with 50 participants to evaluate XAI techniques on GoogLeNet [53], a deep neural network for image classification. Our study generated data for three XAI techniques for this model: LIME [42], Grad-CAM [48], and Feature Visualizations [36]. We identify our core contributions below:

- A **novel, playful GWAP** to **evaluate and compare** the **interpretability** of different XAI techniques and the **agreement between XAI techniques and humans** on image classification tasks.
- We performed an initial **empirical study using a GWAP to evaluate the interpretability of and agreement between XAI techniques and humans** applied to GoogLeNet, a deep neural network for image classification.
- We identify how the data generated by Eye into AI can help researchers **compare the interpretability of different XAI techniques**.

2 RELATED WORK

Using games to collect valid data on crowd-sourcing platforms is a well-established strategy. We provide a brief introduction to the Games With a Purpose literature and how our contribution complements this research area. Furthermore, several XAI techniques have been proposed to provide insight into the AI's prediction. As a result, researchers in human-computer interaction conducted empirical studies to evaluate the interpretability of different techniques and their impact on decision-making. While there are several different XAI techniques, we will only present the three techniques that we used in our game. The three techniques that we chose to use are all different from each other in terms of the methods used to produce the explanation. We will also discuss several empirical studies evaluating XAI techniques.

2.1 Games with a Purpose

Games With a Purpose (GWAPs), also known as Human Computation Games, are a genre of games designed to leverage human computational abilities online [39, 41]. These games make boring tasks more interesting, such as turning image labeling into a playful experience [18, 57]. A GWAP perspective provides benefits over conventional crowd-based tasks. First, GWAPs motivate large-scale participation without relying on financial incentives [10]. In particular, they can activate audiences who might not participate in crowd-work but enjoy game-play [55]. Second, framing an activity as a game improves engagement, leading to higher persistence [26]. Finally, XAI tasks are not a good fit for common data validation strategies, such as similarity to other players' inputs. Instead, thoughtful and creative interpretations are needed. A GWAP framework allows us to take an intrinsic validation approach, where game-play choices are highly aligned with the provision of high-quality human subject data [17]. For our domain of image classification, this means that providing thoughtful responses to ambiguous visual stimuli is the optimal way to play.

From a design perspective, GWAPs involve *action*, *verification*, and *feedback* mechanics [51]. Action mechanics allow the player to solve the human computation problem at hand; verification mechanics collate player data into task-relevant outcomes, including identifying whether the players are providing good data; and feedback mechanics, as the name suggests, provide feedback to players on both their in-game behavior and their task outcomes [50].

GWAPs are varied in their design, and innovation in this space is ongoing. For example, Pe-Than et al. describe a taxonomy of GWAPs that reflects differences in areas such as how the agreement between players is calculated, what incentives are used to foster participation, and whether players can directly interact with one another [39]. Several groups have studied the effects of single- and multi-player GWAPs, as well as competitive and collaborative scoring systems [40, 49]. GWAPs rely on eliciting high-quality player contributions, for example, through intrinsic design methods [17], while weeding out bad data [39]. Explainable AI, however, does not typically consider how to obtain high-quality data from participants [1]. Expanding research on eliciting high-quality data, therefore, expands the fields of both GWAPs and XAI. Additionally, this challenge is particularly important for GWAPs that address XAI, because there is no ground truth other than player understanding. The metrics for evaluating the success of a GWAP include both *task metrics*, such as how many tasks were completed in a given period of time, and *player engagement metrics*, such as how often players return to play again [51]. Success in these metrics is critical for a GWAP, as the games must provide both *good data* and a *good player experience* - otherwise, they have failed to make human computation fun.

2.2 Explainable AI Techniques

Explainable AI (XAI) is commonly defined as making an AI's decision easy to understand by people [13]. Computer vision tasks, such as image classification or object detection, employ deep neural networks (DNNs) that are quite difficult to interpret without XAI techniques. One approach to make black-box models more transparent and interpretable is to develop post-hoc explanations to explain a single prediction, known as a local explanation [34]. There are several types of local explanation techniques to choose from when explaining an individual prediction, such as model-agnostic or pixel attribution techniques.

Local interpretable model-agnostic explanations (LIME) is a model-agnostic method that approximates predictions from black-box models through surrogate, interpretable models [42]. For image data, it shows grouped regions, or superpixels, of an image to highlight the most important superpixels that contributed to the classification of the image. Alternatively, Grad-CAM is a pixel-attribution method that is based on feature maps generated by the last convolutional layer [48]. The resulting saliency map will identify the features in the image that contributed the most to the prediction. As a result, Grad-CAM is visually distinguishable from LIME, as Grad-CAM's saliency maps highlight connected regions by definition, whereas LIME's superpixels are not necessarily contiguous.

Aside from local explanations, there are global explanations that provide insight into the average behavior of the AI [2, 34]. One example of a global explanation is to show the features that the neural network learned through generative examples, known as feature visualizations (FV) [36]. These explanations are abstract representations and do not reveal regions of the original image like LIME and Grad-CAM.

These three XAI techniques are evaluated in our empirical study of Eye into AI as they are distinctly different from one another in terms of methods and visual outputs.

We evaluate local and global techniques to ensure the extensibility of our system for different techniques. The techniques that we evaluate in our empirical study are also highly cited and have open-source implementations that are popular with researchers and practitioners. However, Eye into AI was designed to be extensible to support additional XAI techniques and any image classification model as well.

2.3 Evaluating Explainable AI Techniques

Initial evaluations in human-computer interaction have suggested that explanations have value to users relying on machine learning [52]. Interactive visual analytic systems also offer promise in explaining complex AI by providing interaction techniques to reveal insights about decisions [19]. Various visual systems have been developed to help users understand AI algorithms, such as [20, 23, 38, 60]. However, these evaluations typically focus on task-specific tools for a small number of users. Our proposed GWAP, Eye into AI, intends to provide a verifiable approach to evaluate XAI at scale.

Explainable AI attempts to make AI more transparent and interpretable by attempting to expose the features that led to an AI prediction. As decision-makers increasingly need to understand the models driving the AI [12, 29], there is a new trend towards making AI algorithms fair, accountable, transparent, and interpretable [1]. This trend has become especially popular within the field of human-computer interaction, where researchers have taken a human-centered approach to design and evaluate XAI techniques [14, 15, 22, 25, 32].

2.3.1 Evaluation Metrics. Several studies have designed quantitative metrics based on machine performance to evaluate and compare the "performance", or interpretability, of XAI techniques [27, 28, 45]. For example, Lin et al. measure the performance of multiple local XAI techniques based on "impact score" by generating counterfactuals (i.e., the same image without the most salient region present) and measuring how the absence of the most salient region affects the prediction and confidence of the model [28]. If the absence of the most salient region significantly impacts the model's prediction and confidence, then the XAI technique highlighted an impactful region.

Previously, most works on explainable AI focused "on new algorithms of XAI rather than on usability, practical interpretability and efficacy on real users" [63]. However, in the past few years, numerous studies have focused on evaluating XAI techniques using human-centered methods [4, 22, 30, 33, 59, 62]. Similar to our empirical study, Zhang et al. seek to understand how humans and machines align on an image classification task by evaluating how a post-hoc explanation for three different architectures compares to regions that humans view as most important to the image class [62]. As a result, their main contribution is an understanding of how humans align with the post-hoc explanation for the three architectures they evaluated. Opposite of Zhang et al. [62], we evaluate three different XAI techniques on one architecture in our empirical analysis. Recently, Kim et al. designed a framework to capture how interpretable different XAI techniques are to humans through two metrics: "level of agreement", and "ability to distinguish between correct and incorrect predictions" [22]. This study provides a framework to evaluate four different XAI techniques. H owever, their framework does not use a GWAP approach.

2.3.2 Evaluating XAI through GWAPs. GWAPs have been used for a range of human computation problems, such as generating descriptive labels for music [24], disambiguating words in natural language processing [47], marking segments in text [31], and folding proteins [10]. However, to date, very few works have used a GWAP approach to evaluate and compare XAI techniques. Perhaps the most similar GWAP to Eye Into AI is Peek-a-boom [58], which aims to collect training data for computer vision tasks by having a certain player (Peek) try to guess a word based on parts of an image revealed by a different player (Boom). Instead of revealing parts of an image to help users guess the image like Peek-a-boom [58], our game allows players to reveal and guess using AI explanations to help researchers evaluate XAI techniques. Another GWAP, called FindItOut [5], shares similar challenges and goals to Eye into AI: collecting high-quality data from players to improve downstream AI tasks. Resulting player data from FindItOut can be useful in understanding commonsense question-answering, while player data from Eye into AI can be in understanding XAI techniques. However, very few GWAPs have tackled the problem of explainable AI [39, 51, 56]. Tocchetti et al. propose a GWAP framework that allows researchers to teach non-experts about explainability topics as well as collect data to, “evaluate and enhance the explainability of black-box models” [56]. While the authors’ goals are very similar to those of Eye into AI, their contribution focuses on collecting data to evaluate XAI techniques rather than demonstrating how to use a GWAP to evaluate XAI techniques. One study proposes using a GWAP-like approach to directly compare several different saliency map techniques [30]. They increasingly reveal more portions of an image until the crowd worker can guess the correct image class. While this study is inspired by concepts from “Peek-a-Boom” [58], a GWAP, their work is not designed to be a game like our contribution. Also, unlike our contribution, Lu et al. do not ask crowd-workers to select explanations for a given image to determine the interpretability of a technique [30]. While this study includes a random baseline in the evaluation, a random baseline for each XAI technique is not included. Having a random baseline for each technique can help measure the importance of actual versus random explanations. For example, are players able to guess the correct answer due to the visual presentation of the XAI technique?

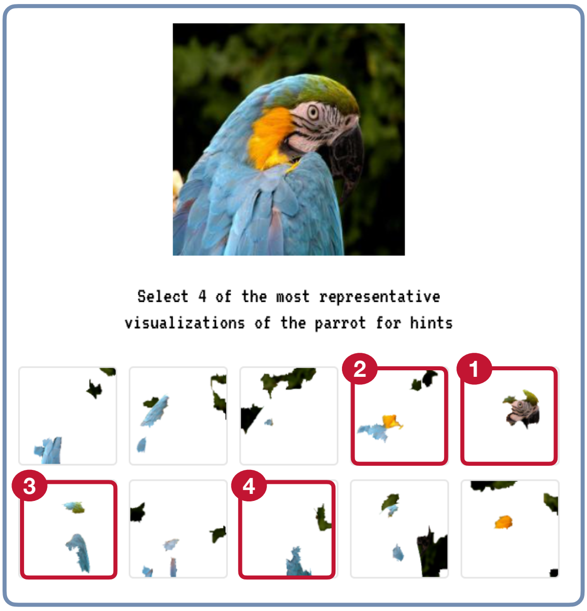
3 DESIGNING EYE INTO AI, A GWAP FOR XAI

We have developed Eye Into AI, an XAI assessment game that focuses on deep learning models and post-hoc XAI techniques for image classification. This task was chosen because interpretability is critical for ensuring reasonable classifications when deployed in real-world settings, as well as for ensuring that there are no harmful biases present in the models. For example, LIME [42], Grad-CAM [48], and Feature Visualizations [36] are popular XAI techniques used by researchers and practitioners to measure how interpretable an image classification model might be. These techniques can yield intuitive, and occasionally beautiful, visual representations that provide clues that the neural network may be behaving properly, making "the hidden layers of networks comprehensible" [7]. However, in practice, such examples of XAI are often handpicked by model builders to demonstrate the efficacy of the network to stakeholders. Therefore, these examples may not provide verifiable evidence that the neural network is performing as expected, as it is difficult to assess the scope and quality of such explanations [3]. Therefore, we identified image classification as a fruitful area for which to design scalable validation techniques with GWAPs.

To develop our prototypes, we used best practices from GWAP design [51] along with Culyba’s Transformational Framework [9]. The latter supports game designers in ensuring their design decisions align with a game’s transformational outcomes. We identified the following as key design goals:

- Players can generate relevant data for explainable AI
- Providing accurate data is the most effective way to play the game
- Data provided by players is copious and timely

To embody these goals in a playable game, we designed a GWAP that involves players taking on multiple roles: an explainer and a guesser.



Select 4 of the most representative visualizations of the parrot for hints

Fig. 2. During the explainer round, the player takes on the role of an explainer and is provided a source image to explain. The player selects four images that they feel best explain the class of the image (i.e., parrot). The explanation shown here represents the XAI technique LIME.

3.1 The Explainer Role

During the explainer round, the player acts as the explainer with the goal of choosing explanations that will help a future player guess the correct answer as fast as possible. The explainer round begins with the game offering the player several categories of images (e.g., instruments, fruits, sea animals) from which they can choose an image that aligns with their interests or curiosities. Upon selecting an image, the explainer is shown ten explanations from an XAI technique. Unbeknownst to the explainer, the best five explanations and worst five explanations (according to the technique) are shown in random order. The goal for the explainer is to create an ordered set of the top four explanations that they believe would allow a future player to correctly guess the original image they selected. For example, in Figure 2, the explainer is shown ten explanations from an XAI technique (LIME), and they have to select the top explanations that they believe will lead other players to guess the correct answer, parrot. The explainer rounds generate quantitative data that can be used to understand how humans agree with the XAI techniques being used.

3.2 The Guesser Role

After the players finish their Explainer round, they will have two rounds as a *Guesser*. One round shows top explanations from an XAI technique, and the other round shows random explanations as a baseline for the same XAI technique. As a guesser, the players attempt to guess the class predicted by the AI as fast as possible. They receive one explanation to start, with a new explanation revealed every ten seconds until they see a total of four visual explanations. The visual explanations for LIME and Grad-CAM are superimposed to build on top of the previous explanations (as seen in Figure 1), while the visual explanations for FV are presented side-by-side in a row. The fewer explanations revealed before a guesser guesses the correct answer, the more points they are awarded. If the guesser is unable to guess the image after all four visual explanations are revealed, they receive a textual hint of the category the image fits in. The textual hints were presented to users in the following format: “It’s a type of [*category*]”, where *category* is one of the seven categories that we coded the image as (refer to Section 4.1). If the guesser is still unable to guess the correct answer, they are shown a multiple choice question listing four possible answers — of which only one is correct. The guesser receives a reduced number of points with these hints. The guesser rounds generate quantitative data that can be used to evaluate the interpretability of the XAI techniques being used.

3.3 Collecting XAI Assessment Data

The game generates relevant data for explainable AI in two ways. First, the explainer is given ten visual explanations to select from, of varying explanatory quality as judged by XAI techniques, but they are only allowed to choose four. They must also select the order in which the images are shown, with the most helpful explanation first. In the instructions for the *explainer* round, players are told to, “Choose features that will be given as hints to other players”.

Second, the text of the guesses provides data on what each guesser hypothesizes each image to be and which explanations helped them guess correctly. Providing accurate data is the best way to play this game because guessers are rewarded with points for getting the right answer. In particular, the design motivates players to get the right answer early so they can receive the maximum number of points. As a result, guessers provide copious and timely data: they can guess as often as they want with no penalty, supporting copiousness, and there is a time limit, keeping their responses timely.

4 EMPIRICAL STUDY: CROWD-SOURCED EVALUATIONS OF EXPLAINABLE AI

Our research explores whether GWAPs can provide meaningful assessments of XAI techniques. To achieve this goal, we deployed Eye Into AI to a) produce an initial dataset from gameplay and b) analyze the dataset to understand the impact of specific XAI explanations on players’ ability to identify what class the AI predicted for the image. In previous research, GWAPs have been shown to be highly effective at collecting data from large crowds. We believe GWAPs can also be effective for assessing XAI techniques embedded within the game, supporting the following hypotheses:

H.1 By collecting self-ranked regions of importance and measuring agreement, we hypothesize that Eye into AI can help researchers quantitatively reveal which techniques humans agree with and which techniques are less intuitive . This information can provide data to help AI researchers improve and optimize their XAI techniques.

H.2 We hypothesize that Eye into AI can help researchers rank the interpretability of techniques quantitatively by measuring exposure and the number of explanations required to lead to a correct interpretation. This will allow researchers to assess when XAI explanations lead to better decisions.

H.3 Finally, we hypothesize that data collected from Eye into AI is scalable even if players have little knowledge of AI or the image classification task at hand by measuring copiousness .

In this experiment, we are only evaluating the interpretability of LIME, Grad-CAM, and FV on one popular deep neural network, GoogLeNet [53], a model for image classification. GoogLeNet is a common architecture in many standard machine learning frameworks and has been identified as a top performer in image classification challenges [44]. Although our study is limited to one model, these results demonstrate how our GWAP can be applied to other popular models and XAI techniques.

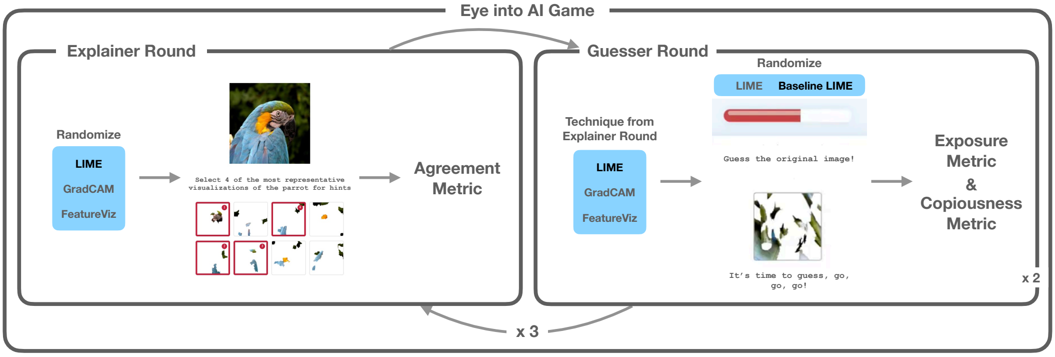


Fig. 3. Design of our empirical study. The left half of the figure shows the explainer round and the resulting metric (i.e., agreement metric) from the data collected for each technique. The right half of the figure shows the guesser round and the resulting metrics from the data (i.e., exposure metric and copiousness metric) collected for each technique and baseline.

Recruitment: In order to have participants play our game, we created a study on Prolific, a crowd-sourcing platform, to serve as a mechanism to compensate participants. We collected results through Prolific from 50 participants (M_{age} = 33.4, SD_{age} = 11.6; 37 female, 13 male) who have English as their first language, currently reside in the United States, have an approval rate of at least 95%, and have a minimum of 50 submissions. Each worker was compensated $1.60 USD for their participation; Prolific reported an average pay of $9.50 USD per hour. Participants were assigned an anonymous ID for analyzing their responses and, on average, took 12 minutes to complete the game. 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 50 participants, 25 participants received bonuses.

Evaluation Metrics: We use three metrics, summarized in Figure 3, to analyze the data generated from players to understand the effectiveness of Eye into AI as an evaluation framework and the effectiveness of different XAI techniques. First, we measure *agreement* between explainers and the XAI techniques, inspired by Kim et al. [22]. This allows us to understand if the explanations that the explainers thought were most useful were also the explanations that the XAI techniques determined as most useful. Second, we measure *exposure*, or how many explanations were needed to be revealed, in order for the guesser to correctly guess the image class. This metric is inspired by a previous crowd-sourcing study for XAI [30]. Exposure allows researchers to rank the interpretability of a given technique. Finally, we measure *copiousness*, or how many guesses a player had, even if they did not end up getting the correct answer.

4.1 Image Selection

For our prototype, we created a dataset of 39 different images based on ImageNet [11] classes that fit into one of seven categories: land animals, sea animals, fruits, vegetables, instruments, transportation, and electronics. Each category had between five to seven images total representing different classes. The images were selected from the Creative Commons collection of Flickr. Each image was selected based on two criteria: (1) there is only one object in the image, and (2) the image has a clear, unambiguous class as determined by the authors. Similar to Zhang et al. [62], we selected images that the AI correctly classified.

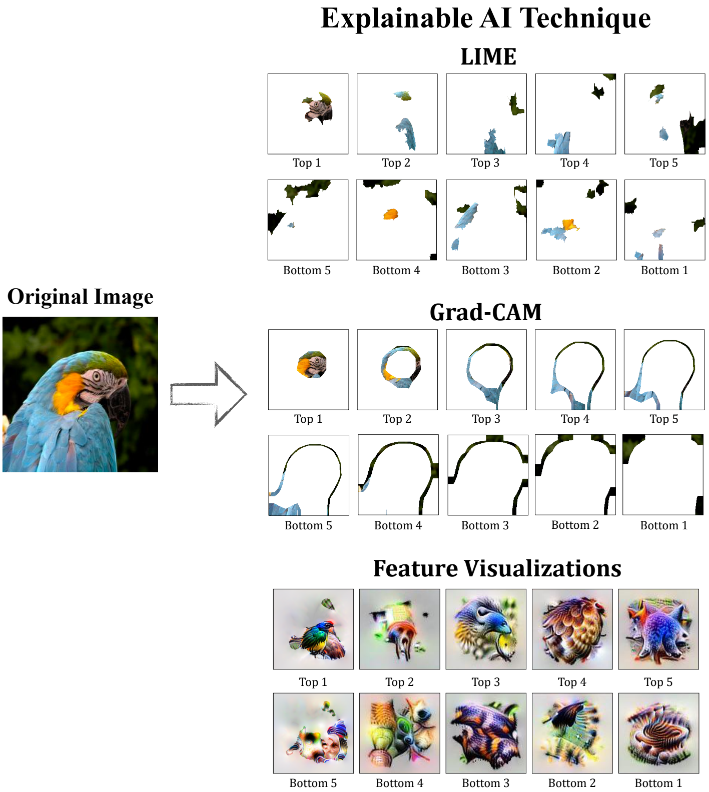


Fig. 4. Example explanations for an image of a parrot. Top shows explanations generated by LIME; the middle shows explanations generated by Grad-CAM; the bottom shows explanations generated by the feature visualization technique.

4.2 Explanation Generation

To create explanations for the game, we used three popular XAI techniques: LIME [43], Grad-CAM [48], and Feature Visualizations (FV) [36]. An example of each explanation for an image of a parrot is shown in Figure 4.

All explanations were generated using the last convolution layer of GoogLeNet, *inception5b* [53]. For LIME and Grad-CAM [16], **we generated one explanation for a single image based on the top prediction for that image**. We partitioned the explanation into ten explanations: the top five explanations and the bottom five explanations. The top five explanations represent the most salient regions of the image, and the bottom five represent the least salient regions of the image. For LIME, the explanations were assigned a rank based on the weight of the superpixel for the prediction. For Grad-CAM, each pixel is given an importance ranking, and we binned pixels based on those rankings (e.g., the top explanation is the top 10% pixels, the second is the top 10-20% of pixels, and so on). This experimental setup ensures that LIME and Grad-CAM reveal the same amount of pixels for each XAI technique.

Unlike LIME and Grad-CAM, which reveal salient portions of the original image, FV algorithmically generates a synthetic image that maximizes a particular neuron [36]. For FV, we used channel attribution [37] to generate explanations for the top five most activated neurons as well

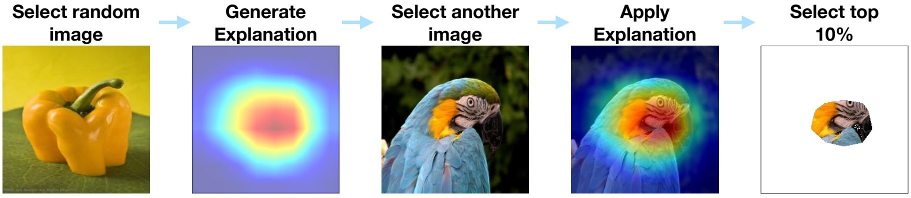


Fig. 5. Pipeline to generate explanations for baseline Grad-CAM. We generated a saliency map for a random image from our dataset and then applied that saliency map on top of another image to generate a random baseline for Grad-CAM.

as the five least activated neurons for the top prediction. The FV explanations were generated using the mixed4d layer of GoogLeNet [54], as neurons in that layer were shown to be semantically meaningful [37]. As FV does not reveal pixels but instead generates a synthetic representation, it is not feasible to constrain by pixels as we did for LIME and Grad-Cam.

We created random baselines for each XAI technique (LIME, Grad-CAM, and FV) to confirm if XAI techniques are better than showing a random explanation. A random baseline is also used in Peek-a-Boom [30], which motivates why we chose to include random baselines in our empirical study. To create the random baselines for LIME, we selected eight random superpixels that were identified for a given image. Five images had fewer superpixels than the other 34 images, so there are only seven random explanations instead of eight for those five images. However, during the game, only four random explanations are selected to be shown out of the seven or eight total random explanations. We ensured that each superpixel represented approximately 10% of all pixels to be consistent with the amount of the image being shown for Grad-CAM, which is why we ended up with seven or eight superpixels in total. To create the random baselines of Grad-CAM, we used the last convolutional layer in GoogLeNet to generate a feature map for a random image from our dataset and applied the resulting saliency map on top of the primary image. For example, in Figure 5, the heatmap of an image of a bell pepper was masked on top of an image of a parrot to get the top and bottom explanations.

To create the FV random baselines, we randomly selected ten of the 528 neuron explanations from layer mixed4d. We excluded explanations that were algorithmically ranked by channel attribution [37] as the top five explanations and the bottom five explanations for that image.

4.3 Game Flow

Participants played the game three times, one for each XAI technique. The order of techniques across games was assigned randomly. Each game began with one explainer round and two guesser rounds. Of the two guesser rounds, they are also randomly ordered: one features the top results of that game's XAI technique, and the other features the corresponding random baseline. After the participant plays through the game all three times, they are asked to complete a survey with nine questions.

5 EMPIRICAL STUDY RESULTS

We present the results of our empirical study evaluating the interpretability of the three XAI techniques with GoogLeNet [53] using Eye into AI. This empirical study was conducted to gather an initial dataset of results from the game to answer our three hypotheses. We note that this particular study only used images that the model correctly classified, but the game is extensible to capture data for when the AI is incorrect as well. The results presented below are only described to

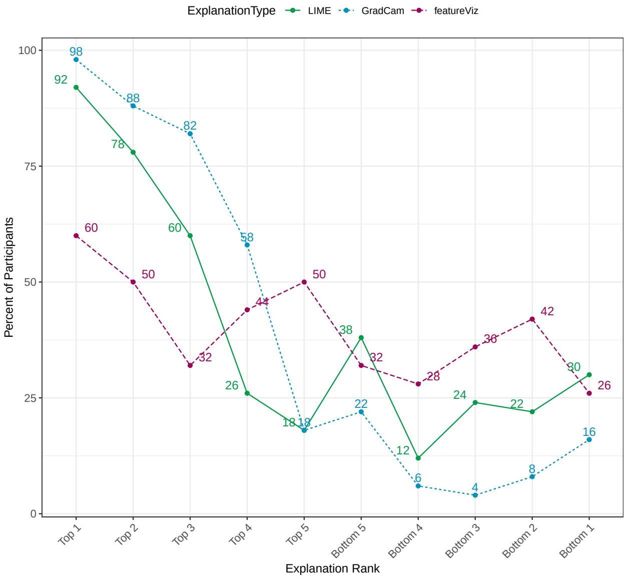


Fig. 6. Measuring *agreement* between the explainer and the XAI technique. The percentage of participants who selected each type of explanation as one of their four selections.

demonstrate the capabilities of Eye into AI. We identify how the data from our empirical study support our three hypotheses.

5.1 Hypothesis 1: Measuring agreement between humans and XAI

As described in Section 3.1, players select, in order, the top four explanations that they think are the most representative of the image (e.g., Figure 2). To determine if and how Eye into AI can help researchers identify the agreement between humans and XAI techniques (**H.1**), we measure agreement between *explainers* and the XAI techniques. More specifically, we analyze the number of players who selected an explanation ranked high or low by LIME, Grad-CAM, and FV. To determine the statistical significance of a participant selecting a top-ranked versus a bottom-ranked explanation, we aggregated the ten explanations to fit into two categories: top or bottom. Within these two categories, top and bottom, we represent how many explanations in each category a player selected. We utilize a Mann-Whitney U test with Benjamini-Hochberg corrections to determine statistical significance as the results do not have a normal distribution.

As seen in Figure 6, we observed that the players had a higher level of agreement for the top-ranked explanations with LIME and Grad-CAM than FV. 86% of the selected explanations for Grad-CAM were top-ranked explanations, while only 14% of the selected explanations were bottom-ranked explanations ($p < 0.001$). For LIME, 68.5% of the selected explanations were top-ranked explanations, while 31.5% of the selected explanations were bottom-ranked explanations ($p < 0.001$).

The results from the participants suggest that this task was more challenging for FV. As one participant remarked about FV, "...the [explanations] were nothing like what the object was. The [explanations] for strawberry were colorful paintings of animals, for instance." Indeed, only 59% of the explanations selected by players for FV were top-ranked explanations. Less than half (41%) of the selected explanations were bottom-ranked explanations. Nonetheless, a Mann-Whitney U test shows that there is a statistically significant difference between the number of top-ranked explanations selected versus bottom-ranked explanations for FV ($p < 0.001$), which showed players could clearly differentiate these groupings.

By gathering quantitative data generated from playing Eye into AI, researchers can use the *agreement* metric to measure and compare the agreement between humans and various XAI techniques through statistical tests and visualizations, which supports **H.1** and provides evidence that researchers can use Eye into AI for this type of analysis (**RQ1**).

5.2 Hypothesis 2: Measuring exposure to rank the interpretability of XAI techniques

We measure interpretability using Kim et al.'s definition that "a method is interpretable if a user can correctly and efficiently predict the method's results" [21]. Within the guesser round, players were shown up to four explanations to help them determine the predicted class of the image that was revealed to them. Every player did this for each XAI technique (i.e., LIME, Grad-CAM, and FV) and each baseline technique. To determine if and how Eye into AI can help researchers rank the interpretability of techniques (**H.2**), we measure *exposure* of explanations. Specifically, we measured the number of participants that guessed correctly when shown a certain number of explanations. The data for these analyses meet the following conditions: (1) the distribution of the data is not normal, (2) it follows a Likert or ordinal scale, (3) the data do not have equal variances, and (4) the observations are independently sampled. In this case, we performed a Kruskal-Wallis test and Dunn's post-hoc test using the Benjamini–Hochberg correction method to determine the statistical significance of the differences among the explanation techniques.

5.2.1 Measuring exposure allows for the comparison of interpretability techniques. The top explanation generated by LIME and Grad-CAM aims to represent the most salient region of the image for the model's prediction. In Figure 7, 54% of the participants were able to guess correctly when shown the top explanation generated by Grad-CAM, compared to 36% with LIME. Recall that LIME and Grad-CAM are visually distinct and algorithmically different from one another, resulting in different pixels for the same image being shown for the explanation. Dunn's post-hoc test shows that this difference observed between LIME and Grad-CAM is statistically significant ($p < 0.05$).

Grad-CAM and LIME are statistically significantly different from FV and all three baselines in Table 1. These results suggest that Grad-CAM is significantly better than LIME and FV. Further, LIME is significantly better than FV and the baselines. Players agreed in the post-game survey, as one participant shared, "The one that shows bits of the photos themselves [were easiest]."

Players performed significantly worse on random baselines for LIME and Grad-CAM in terms of correctly guessing the image when shown only the first explanation. Dunn's test shows statistical significance between LIME and baseline LIME ($p < 0.001$) and between Grad-CAM and baseline Grad-CAM ($p < 0.001$). Figure 7 shows that only 8% of the players were able to get the correct answer when shown only the first explanation for the baseline Grad-CAM while 54% of the players were able to get the correct answer when shown the first explanation for Grad-CAM. Similarly, only 6% of the players got the correct answer when shown the first explanation for baseline LIME, while 36% of the players got the correct answer when shown LIME.

Table 1 shows that Grad-CAM is significantly better than LIME, FV, and all baselines after a single explanation. Further, LIME is significantly better than FV and the baselines. Notably, FV does not perform significantly better than any random baselines.

While the players did a decent job at correctly guessing the image class when shown LIME and Grad-CAM, they struggled to determine the image class when shown FV explanations. Dunn's post-hoc test (Table 2) shows there is a statistically significant difference between the number of participants who guessed correctly when shown up to four explanations generated by LIME versus

Table 1. Dunn's post-hoc test with the Benjamini-Hochberg correction method for determining statistical significance between XAI techniques when only shown first explanation.

Comparison	P.unadj	P.adj
Grad-CAM - LIME	0.02	0.03
Grad-CAM - B_LIME	<0.001	<0.001
Grad-CAM - FV	<0.001	<0.001
Grad-CAM - B_FV	<0.001	<0.001
Grad-CAM - B_Grad-CAM	<0.001	<0.001
LIME - FV	<0.001	<0.001
LIME - B_FV	<0.001	<0.001
LIME - B_LIME	<0.001	<0.001
LIME - B_Grad-CAM	<0.001	<0.001

when shown explanations generated by FV ($p < 0.001$). There is also a statistically significant difference between the number of participants who guessed correctly when shown explanations generated by Grad-CAM versus when shown explanations generated by FV ($p < 0.001$). As one participant remarked, "The abstract [FV explanations] were almost impossible to guess. They were too abstract."

As seen in Table 2, there was no statistical significance between the performance of the guessers when shown explanations for FV and the random baseline of FV ($p = 0.09$). However, it should be noted that only one participant out of 50 guessed correctly based on the random baseline, whereas ten participants out of 50 guessed correctly when shown the explanations for FV (Figure 7). We suggest future work to validate whether a larger participant pool would result in a statistically significant difference between the top explanations for FV and the random baseline.

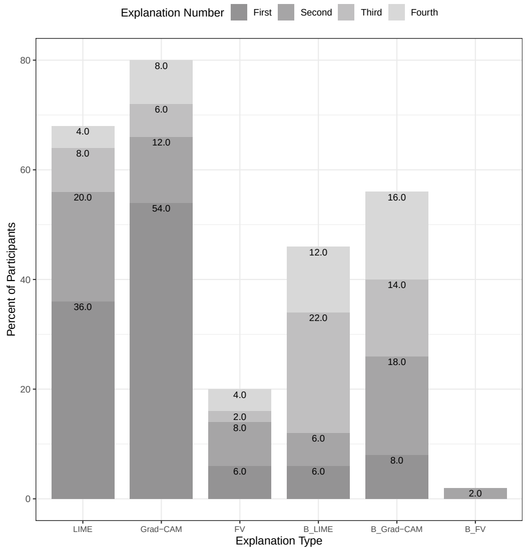


Fig. 7. Percent of participants that correctly guess the image when shown the first, second, third, or fourth explanation for a given explanation technique.

Table 2. Dunn's post-hoc test with the Benjamini-Hochberg correction method for determining statistical significance between XAI techniques when shown 1-4 explanations.

Comparison	P.unadj	P.adj
Grad-CAM - LIME	0.23	0.26
Grad-CAM - B_LIME	<0.001	<0.001
Grad-CAM - FV	<0.001	<0.001
Grad-CAM - B_FV	<0.001	<0.001
Grad-CAM - B_Grad-CAM	0.02	0.02
LIME - FV	<0.001	<0.001
LIME - B_FV	<0.001	<0.001
LIME - B_LIME	0.03	0.04
LIME - B_Grad-CAM	0.23	0.25
FV- B_FV	0.07	0.09
FV- B_Grad-CAM	<0.001	<0.001
FV- B_LIME	0.01	0.02

By gathering quantitative data generated from playing Eye into AI, we observed that researchers could use the *exposure* metric to measure and compare interpretability techniques through statistical tests and visualizations, which supports **H.2** and answers **RQ2**.

5.3 Hypothesis 3: Eye into AI is a scalable XAI evaluation method

After the participants played three rounds of Eye into AI, they took a survey with several questions on a 5-point Likert scale (strongly disagree, disagree, neutral, agree, and strongly agree), as well as a few open-ended questions. Our goal was to capture details about their game experience and any prior knowledge of AI. We only report results from the survey that we believe are most beneficial to understanding the game and the players' perceptions of it[1].

From the survey, we found that 70% of the players do not consider themselves knowledgeable about AI. Nonetheless, 80% of the players thought it was clear how to play the game. These findings support our third hypothesis that an XAI GWAP can be played by non-experts, which in turn supports the scalability of our method.

We also asked players if they would like to keep playing the game, a proxy question that game designers use to measure enjoyment [46]. 70% of the players agreed or strongly agreed with this statement. Finally, players had the opportunity to optionally provide any other comments they had about the game at the end of the survey. Out of the 50 players, 27 responded to this question, which included statements such as "The game was fun, engaging, and made me think", "It was fun! Especially when those weird [FV explanations] appeared." and "I loved it, can I play it now? You should totally make an app, I would play all the time!".

Additionally, players were copious with coming up with creative and thoughtful guesses, even if the XAI technique did not lead them to the correct guess. As Figure 8 illustrates, players contributed numerous guesses during each round, even for the more difficult representations. We observed that players had more guesses for FV and baseline FV compared to the LIME and Grad-CAM and their baselines. For example, even though FV explanations only led to 20% accuracy (as shown in Figure 7), players on average contributed more than four guesses. This is expected as players often did not uncover the correct answer for FV and baseline FV and kept guessing. However, this data

[1] View the supplemental material to see all the post-game survey questions.

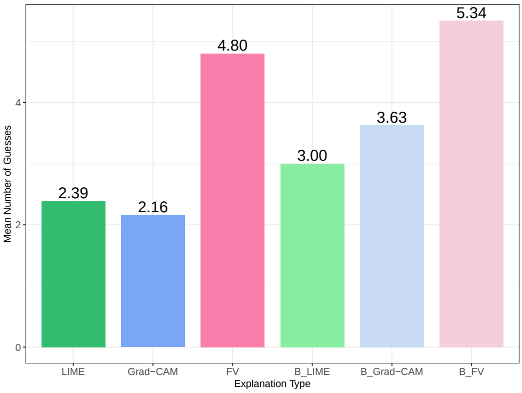


Fig. 8. Mean number of guesses from participants during each guesser round.

has the opportunity to be further analyzed by XAI researchers to see how explanations may be misinterpreted by real people.

6 DISCUSSION

We created a Game with a Purpose called Eye into AI to aid researchers in evaluating XAI techniques. The game features two distinct rounds: the guesser role and the explainer role. The explainer role gathers data to analyze the agreement between humans and XAI techniques using the *agreement* metric, while the guesser role gathers data to analyze the interpretability of a technique using the *exposure* metric. We conducted an initial empirical study on Eye into AI to understand whether Eye into AI could produce valid data for researchers to use. We used three metrics (agreement, exposure, and copiousness) and qualitative survey data to understand how Eye into AI can support various analyses that researchers may use to evaluate and compare XAI techniques. We briefly discuss our findings and how they can be helpful to XAI researchers below.

During the explainer round, we asked players to select four out of ten explanations generated by an XAI technique that they thought best represented the main object in the image. We found that using the agreement metric, we were able to quantify the agreement between humans and each XAI technique. Through statistical tests, we were able to rigorously compare the three techniques. The ability of humans to distinguish between the top- and bottom-ranked explanations aligns with cognitive science principles of visual attention. Specifically, bottom-up visual attention is a process in which humans shift their attention to the salient features of images without activating more complex cognitive processes [8]. Being able to distinguish between explanations that are top-ranked versus bottom-ranked explanations provides researchers insight into which XAI techniques align with humans.

One of the architectures Zhang et al. [62] compared in their experiments was an Inception V3, which is based on the inception module and is used in the GoogLeNet architecture [53]. Given that the architecture we used is very similar to one of the architectures used by Zhang et al., we briefly discuss how our findings compare. Although we did not include SHAP in our experiment, we see that other saliency map techniques (e.g., LIME and Grad-CAM) show high agreement between the players and GoogLeNet, which aligns with the trends observed by Zhang et al. on Inception V3 with SHAP [62].

During the guesser round, we revealed explanations to the players to help them guess the predicted class. We showed the players three XAI techniques and their corresponding baselines during the guessing round. By using the exposure metric, researchers can compare the interpretability of different techniques. Zhang et al. also measure exposure, albeit slightly differently than we do, and observe that very few images were correctly recognized by humans in the first five segments for Inception V3. We observed that more than half of the participants were able to correctly identify the image when shown only the first explanation of Grad-CAM, and less than half of the participants were able to correctly identify the image when shown only the first explanation of FV. While LIME and SHAP are two very different methods, it raises interesting questions about how Grad-CAM is different from these two methods, which allowed more participants to correctly recognize the predicted class for this model. It also begs the question of to what extent there is an interaction effect between the model and the XAI technique. Indeed, Zhang et al. also saw different performances for different model architectures trained on the same data.

Existing GWAPs like Peek-a-boom [58] were not designed to evaluate XAI techniques, and the experiments conducted by previous works evaluating XAI techniques were not explicitly designed to be fun games. With Eye into AI, we integrate existing metrics into an enjoyable game that collects valid data at scale for researchers. We observed that 70% of the players would like to continue playing the game, which we used as a proxy for enjoyment. However, Eye into AI is just one way to evaluate XAI techniques. Eye into AI demonstrates to XAI researchers that games are an efficient way to capture human-centered data to evaluate their techniques.

7 LIMITATIONS AND FUTURE WORK

In this section, we briefly address limitations that stem from the design of Eye into AI as well as the empirical study of Eye into AI. By addressing limitations with our work, we identify several avenues that future work should consider.

7.1 Lack of multiplayer functionality.

Eye into AI was originally conceived to have multiplayer functionality. Although this has yet to be implemented, with this feature, the game would allow further analyses on how the explanations that explainers rank directly impact the ability of other players, such as *guessers* to guess correctly, similar to Zhang et al. [62]. With the lack of multiplayer functionality, we can only compare how explainers rank explanations to how XAI techniques do so. Future work should develop this multi-player feature and analyze how explanations players choose, compared to XAI explanations, impact another player's ability to guess correctly in a game setting.

7.2 Lack of analysis on guesses throughout a round

Eye into AI generates a lot of data, including how many guesses a player makes each round and what they guess each time. These data capture how the player understands the explanations throughout the round and how they may be misinterpreting the explanation. For example, when a player was shown the top four LIME explanations for broccoli, the player guessed lettuce and spinach. By analyzing the guesses, researchers may gain insight as to why explanations are not as effective as they could be. For example, it can help researchers identify when a model may be relying on spurious patterns. Future work should explore analyzing how the guesses evolve throughout each round and how this can help researchers improve XAI.

7.3 Aimed to explore correct classifications only

In our empirical study of Eye into AI, our goal was to explore the system for images that the model correctly classified. This limits our analyses because we cannot interpret how the players viewed the explanations when the classification was incorrect. Furthermore, when measuring the interpretability of XAI techniques, it is necessary to also explore images that were incorrectly classified. We encourage future studies using Eye Into AI to include images that the model incorrectly classifies to better represent the interpretability of a technique. It is unclear to what extent the explainer round instructions to "guess the original image" would have significantly impacted our results in Section 5.2, since the images were correctly classified by the model. In future versions of Eye into AI, users should receive explicit instructions during the guessing round to "guess which class the AI predicted for this image".

7.4 Empirical analysis limited to one model and three XAI techniques.

We only empirically tested our GWAP with one model and three XAI techniques. Future work is encouraged to explore the same techniques on different models for the same domain in order to understand to what extent model design and performance impact the interpretability of different XAI techniques. Eye into AI can be easily configured to evaluate the same interpretability techniques used on different architectures, different interpretability techniques used on different architectures, and different interpretability techniques used on one architecture (e.g., our empirical analysis presented in this paper).

7.5 No clear guidelines for generating random baselines for XAI techniques

We created random baselines for each explanation technique to better understand if formalized XAI techniques are more or less interpretable than a random baseline. Since there is no defined way to make random baselines, we created our own techniques. While we tried to align it with how each XAI technique was designed, we acknowledge that this may impact our analysis.

8 CONCLUSION

With several XAI techniques being developed to improve the transparency and interpretability of AI, being able to evaluate and compare them has become increasingly important. Several XAI evaluation studies have used machine-based performance to identify which XAI techniques are more interpretable than others, but these methods may not generalize to how humans view these techniques. While recent studies use crowd-sourcing and human-based metrics to evaluate XAI techniques, we develop a GWAP, Eye into AI, and conduct an empirical study to determine if Eye into AI is a valid XAI evaluation method to collect data at scale.

We designed Eye into AI, a GWAP, to help researchers collect data to evaluate and compare the interpretability of visually distinct and algorithmically different XAI techniques. Through an empirical study, we evaluated how well our GWAP achieves that goal by exploring the significance of player data using three metrics. Our novel, playful method for evaluating and comparing XAI techniques produced statistically significant results regarding how humans agree or disagree with a technique's ranking of explanations by measuring *agreement* and which techniques may be more interpretable than others by measuring *exposure* and *task accuracy*. Furthermore, we found that the players provide numerous guesses by measuring *copiousness*, and they do not need to have knowledge about AI in order to generate valid data while playing Eye into AI. Evaluating explainable AI techniques through games with a purpose, such as Eye into AI, ultimately can provide researchers with new methods to investigate how humans perceive and collaborate with AI on visual decision-making tasks.

ACKNOWLEDGMENTS

We would like to thank everyone who played the game to help us iterate on the design.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. ACM, New York, NY, USA, Article 582, 18 pages. https://doi.org/10.1145/3173574.3174156

[2] 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

[3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps. In *Proceedings of the 32Nd International Conference on Neural Information Processing Systems* (Montreal, Canada) *(NIPS'18)*. Curran Associates Inc., USA, 9525–9536. http://dl.acm.org/citation.cfm?id=3327546.3327621

[4] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study. *CoRR* abs/2002.00772 (2020). arXiv:2002.00772 https://arxiv.org/abs/2002.00772

[5] Agathe Balayn, Gaole He, Andrea Hu, Jie Yang, and Ujwal Gadiraju. 2022. Ready Player One! Eliciting Diverse Knowledge Using A Configurable Game. In *Proceedings of the ACM Web Conference 2022*. 1709–1719. https://doi.org/10.1145/3485447.3512241

[6] 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

[7] Shan Carter, Zan Armstrong, Ludwig Schubert, Ian Johnson, and Chris Olah. 2019. Activation Atlas. *Distill* (2019). https://distill.pub/2019/activation-atlas

[8] Charles E. Connor, Howard E. Egeth, and Steven Yantis. 2004. Visual Attention: Bottom-Up Versus Top-Down. *Current Biology* 14, 19 (Oct 2004), R850–R852. https://doi.org/10.1016/j.cub.2004.09.041

[9] Sabrina Culyba. 2018. *The Transformational Framework: A process tool for the development of Transformational games.* figshare. https://kilthub.cmu.edu/articles/journal_contribution/The_Transformational_Framework_A_Process_Tool_for_the_Development_of_Transformational_Games/7130594/files/13117568.pdf

[10] Vickie Curtis. 2015. Motivation to participate in an online citizen science game: A study of Foldit. *Science Communication* 37, 6 (2015), 723–746. https://doi.org/10.1177/1075547015609322

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255. https://doi.org/10.1109/cvpr.2009.5206848

[12] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. arXiv:1702.08608 [stat.ML] https://arxiv.org/pdf/1702.08608

[13] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: Towards social transparency in ai systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19. https://doi.org/10.1145/3411764.3445188

[14] Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable ai: Towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*. Springer, 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. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–6. https://doi.org/10.1145/3411763.3441342

[16] Jacob Gildenblat and contributors. 2021. PyTorch library for CAM methods. https://github.com/jacobgil/pytorch-grad-cam.

[17] David Gundry and Sebastian Deterding. 2018. Intrinsic elicitation: A model and design approach for games collecting human subject data. In *Proceedings of the 13th International Conference on the Foundations of Digital Games*. ACM, 38. https://doi.org/10.1145/3235765.3235803

[18] Chien-Ju Ho, Tao-Hsuan Chang, Jong-Chuan Lee, Jane Yung-jen Hsu, and Kuan-Ta Chen. 2010. KissKissBan: A Competitive Human Computation Game for Image Annotation. *SIGKDD Explor. Newsl.* 12, 1 (Nov. 2010), 21–24. http://doi.acm.org/10.1145/1882471.1882475

[19] Fred Matthew Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. 2018. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE transactions on visualization and computer graphics* (2018). https://doi.org/10.1109/tvcg.2018.2843369

[20] Matthew Kay, Tara Kola, Jessica R Hullman, and Sean A Munson. 2016. When (ish) is my bus?: User-centered visualizations of uncertainty in everyday, mobile predictive systems. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5092–5103. https://doi.org/10.1145/2858036.2858558

[21] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems* 29 (2016). https://proceedings.neurips.cc/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf

[22] Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. 2022. HIVE: Evaluating the Human Interpretability of Visual Explanations. *arXiv:2112.03184 [cs]* (Jan 2022). https://doi.org/10.1007/978-3-031-19775-8_17 arXiv: 2112.03184.

[23] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. ACM, New York, NY, USA, 5686–5697. http://doi.acm.org/10.1145/2858036.2858529

[24] Edith LM Law, Luis Von Ahn, Roger B Dannenberg, and Mike Crawford. 2007. TagATune: A Game for Music and Sound Annotation.. In *ISMIR*, Vol. 3. 2. https://www.cs.cmu.edu/~elaw/papers/ISMIR2007.pdf

[25] Q Vera Liao and Kush R Varshney. 2021. Human-centered explainable AI (XAI): From algorithms to user experiences. *arXiv preprint arXiv:2110.10790* (2021). https://arxiv.org/pdf/2110.10790

[26] Andreas Lieberoth. 2015. Shallow gamification: Testing psychological effects of framing an activity as a game. *Games and Culture* 10, 3 (2015), 229–248. https://doi.org/10.1177/1555412014559978

[27] Yi-Shan Lin, Wen-Chuan Lee, and Z. Berkay Celik. 2021. What Do You See?: Evaluation of Explainable Artificial Intelligence (XAI) Interpretability through Neural Backdoors. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (2021). https://doi.org/10.1145/3447548.3467213

[28] 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). https://arxiv.org/pdf/1910.07387.pdf

[29] Zachary Chase Lipton. 2016. The Mythos of Model Interpretability. *CoRR* abs/1606.03490 (2016). arXiv:1606.03490 http://arxiv.org/abs/1606.03490

[30] Xiaotian Lu, Arseny Tolmachev, Tatsuya Yamamoto, Koh Takeuchi, Seiji Okajima, Tomoyoshi Takebayashi, Koji Maruhashi, and Hisashi Kashima. 2021. Crowdsourcing Evaluation of Saliency-based XAI Methods. *arXiv:2107.00456 [cs]* (Aug 2021). http://arxiv.org/abs/2107.00456 arXiv: 2107.00456.

[31] Chris Madge, Jon Chamberlain, Udo Kruschwitz, and Massimo Poesio. 2017. Experiment-driven development of a gwap for marking segments in text. In *Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play*. ACM, 397–404. https://doi.org/10.1145/3130859.3131332

[32] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[33] Sina Mohseni and Eric D. Ragan. 2018. A Human-Grounded Evaluation Benchmark for Local Explanations of Machine Learning. *CoRR* abs/1801.05075 (2018). arXiv:1801.05075 http://arxiv.org/abs/1801.05075

[34] Christoph Molnar. 2019. *Model-Agnostic Methods*. Lulu. https://christophm.github.io/interpretable-ml-book/agnostic.html

[35] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D Ragan. 2019. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 97–105. https://doi.org/10.1609/hcomp.v7i1.5284

[36] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. 2017. Feature Visualization. *Distill* (2017). https://distill.pub/2017/feature-visualization

[37] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. 2018. The Building Blocks of Interpretability. *Distill* (2018). https://distill.pub/2018/building-blocks

[38] Kayur Patel, Naomi Bancroft, Steven M Drucker, James Fogarty, Andrew J Ko, and James Landay. 2010. Gestalt: integrated support for implementation and analysis in machine learning. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*. ACM, 37–46. https://doi.org/10.1145/1866029.1866038

[39] Ei Pa Pa Pe-Than, Dion Hoe-Lian Goh, and Chei Sian Lee. 2015. A typology of human computation games: an analysis and a review of current games. *Behaviour & Information Technology* 34, 8 (2015), 809–824. https://doi.org/10.1080/0144929x.2013.862304

[40] Ei Pa Pa Pe-Than, Dion Hoe-Lian Goh, and Chei Sian Lee. 2017. Does it matter how you play? The effects of collaboration and competition among players of human computation games. *Journal of the Association for Information Science and Technology* 68, 8 (2017), 1823–1835. https://doi.org/10.1002/asi.23863

[41] Alexander J. Quinn and Benjamin B. Bederson. 2011. Human Computation: A Survey and Taxonomy of a Growing Field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. ACM, New York, NY, USA, 1403–1412. https://doi.org/10.1145/1978942.1979148

[42] 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

[43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. (Aug 2016), 1135–1144. https://doi.org/10.1145/2939672.2939778

[44] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

[45] Sam Sattarzadeh, Mahesh Sudhakar, and Konstantinos N. Plataniotis. 2021. SVEA: A Small-scale Benchmark for Validating the Usability of Post-hoc Explainable AI Solutions in Image and Signal Recognition. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (2021), 4141–4150. https://doi.org/10.1109/iccvw54120.2021.00462

[46] Henrik Schoenau-Fog et al. 2011. The Player Engagement Process-An Exploration of Continuation Desire in Digital Games.. In *Digra conference*. [PDF]digra.org

[47] Nitin Seemakurty, Jonathan Chu, Luis Von Ahn, and Anthony Tomasic. 2010. Word sense disambiguation via human computation. In *Proceedings of the acm sigkdd workshop on human computation*. ACM, 60–63. https://doi.org/10.1145/1837885.1837905

[48] 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

[49] Kristin Siu, Matthew Guzdial, and Mark O. Riedl. 2017. Evaluating Singleplayer and Multiplayer in Human Computation Games. In *Proceedings of the 12th International Conference on the Foundations of Digital Games* (Hyannis, Massachusetts) *(FDG '17)*. ACM, New York, NY, USA, Article 34, 10 pages. http://doi.acm.org/10.1145/3102071.3102077

[50] Kristin Siu and Mark O. Riedl. 2016. Reward Systems in Human Computation Games. In *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play* (Austin, Texas, USA) *(CHI PLAY '16)*. ACM, New York, NY, USA, 266–275. http://doi.acm.org/10.1145/2967934.2968083

[51] Kristin Siu, Alexander Zook, and Mark O. Riedl. 2017. A Framework for Exploring and Evaluating Mechanics in Human Computation Games. In *Proceedings of the 12th International Conference on the Foundations of Digital Games* (Hyannis, Massachusetts) *(FDG '17)*. ACM, New York, NY, USA, Article 38, 4 pages. http://doi.acm.org/10.1145/3102071.3106344

[52] Simone Stumpf, Vidya Rajaram, Lida Li, Weng-Keen Wong, Margaret Burnett, Thomas Dietterich, Erin Sullivan, and Jonathan Herlocker. 2009. Interacting Meaningfully with Machine Learning Systems: Three Experiments. *Int. J. Hum.-Comput. Stud.* 67, 8 (Aug. 2009), 639–662. http://dx.doi.org/10.1016/j.ijhcs.2009.03.004

[53] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2014. Going Deeper with Convolutions. *arXiv:1409.4842 [cs]* (Sep 2014). http://arxiv.org/abs/1409.4842 arXiv: 1409.4842.

[54] Tensorflow. 2020. Lucid. https://github.com/tensorflow/lucid.

[55] Stefan Thaler, Elena Simperl, and Stephan Wölger. 2012. An experiment in comparing human-computation techniques. *IEEE Internet Computing* 16, 5 (2012), 52–58. https://doi.org/10.1109/mic.2012.67

[56] Andrea Tocchetti, Marco Brambilla, Lorenzo Corti, and Irene Celino. 2022. EXP-Crowd: a Gamified Crowdsourcing Framework for Explainability. *Frontiers in Artificial Intelligence* (2022), 61. https://doi.org/10.3389/frai.2022.826499

[57] Luis Von Ahn and Laura Dabbish. 2008. Designing games with a purpose. *Commun. ACM* 51, 8 (2008), 58–67. https://dl.acm.org/doi/fullHtml/10.1145/1378704.1378719

[58] Luis von Ahn, Ruoran Liu, and Manuel Blum. 2006. Peekaboom: a game for locating objects in images. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. Association for Computing Machinery, 55–64. https://doi.org/10.1145/1124772.1124782

[59] 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, New York, NY, USA, 318–328. https://doi.org/10.1145/3397481.3450650

[60] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE transactions on visualization and computer graphics* (2019). https://doi.org/10.1109/tvcg.2019.2934619

[61] 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

[62] Zijian Zhang, Jaspreet Singh, Ujwal Gadiraju, and Avishek Anand. 2019. Dissonance between human and machine understanding. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–23. https://doi.org/10.1145/3359158

[63] Jichen Zhu, Antonios Liapis, Sebastian Risi, Rafael Bidarra, and G Michael Youngblood. 2018. Explainable AI for designers: A human-centered perspective on mixed-initiative co-creation. In *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 1–8. https://doi.org/10.1109/cig.2018.8490433

Received July 2022; revised January 2023; accepted April 2023