# Getting Playful with Explainable AI: Games with a Purpose to Improve Human Understanding of AI

**Laura Beth Fulton**
Carnegie Mellon University
Pittsburgh, PA 15213, USA
lfulton@andrew.cmu.edu

**Ja Young Lee**
Carnegie Mellon University
Pittsburgh, PA 15213, USA
haeza37@gmail.com

**Qian Wang**
Carnegie Mellon University
Pittsburgh, PA 15213, USA
wangqian.evelyn@gmail.com

**Zhendong Yuan**
Carnegie Mellon University
Pittsburgh, PA 15213, USA
zhendony@andrew.cmu.edu

**Jessica Hammer**
Carnegie Mellon University
Pittsburgh, PA 15213, USA
hammerj@andrew.cmu.edu

**Adam Perer**
Carnegie Mellon University
Pittsburgh, PA 15213, USA
adamperer@cmu.edu


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

## Abstract

Explainable Artificial Intelligence (XAI) is an emerging topic in Machine Learning (ML) that aims to give humans visibility into how AI systems make decisions. XAI is increasingly important in bringing transparency to fields such as medicine and criminal justice where AI informs high consequence decisions. While many XAI techniques have been proposed, few have been evaluated beyond anecdotal evidence. Our research offers a novel approach to assess how humans interpret AI explanations; we explore this by integrating XAI with Games with a Purpose (GWAP). XAI requires human evaluation at scale, and GWAP can be used for XAI tasks which are presented through rounds of play. This paper outlines the benefits of GWAP for XAI and demonstrates application through our creation of a multi-player GWAP which focuses on explaining deep learning models trained for image recognition. Through our game, we seek to understand how humans select and interpret explanations used in image recognition systems and bring empirical evidence on the validity of GWAP designs for XAI.

## Author Keywords

Explainable AI, Games with a Purpose, Interpretable Machine Learning, Visualization

## Introduction

Artificial Intelligence (AI) is increasingly becoming ubiquitous in day-to-day human experiences. There are claims that AI has exceeded human performance in certain domains, and its use has proliferated in fields such as healthcare and criminal justice where AI informs high consequence decisions. However, human decision making is reliant on rational which is driven by forms of observation, intuition, experience, and logical thinking [13]. Complex AI systems lack the ability to self-explain their thought processes in ways which humans can interpret. The future is reliant on human trust and understanding of models which can be supported through the emerging field of Explainable AI (XAI). We believe making models explainable is a prerequisite for building trust and understanding AI systems at scale.

Games with a Purpose (GWAP) are games designed to generate usable data as a byproduct of gameplay. These games are designed to make boring tasks, such as labeling data, more interesting [1, 14]. GWAP have been shown to be highly effective at collecting and validating large sets of data generated from online gameplay [1]. GWAP have been used for numerous human computation problems, such as generating descriptive keywords for music [17], labeling speech information for natural language processing [23], making segments in text [6], and folding proteins [2]. We apply GWAP to tackle the problem of XAI, which we believe is the first GWAP designed with this focus [4, 26].

XAI is an appropriate fit for GWAP as it requires human evaluation for tasks that can be split into smaller parts. It is typical for GWAP to rely on collecting large quantities of data and then exclude low-quality data. Common validation strategies (e.g. agreement design where player contributions are evaluated based on similarity to the majority opinion shared by other players) are not appropriate for explainability tasks [7]. For explainability, thoughtful and creative interpretations are needed to assess if explanations are sensible.

Through GWAP for XAI design, we focus on an intrinsic validation approach, where gameplay choices are aligned such that they produce data that represents players' creative interpretations. Our contributions include providing a verifiable approach to evaluate XAI through GWAP design and experimentation.

In summary, our contributions for this research include 1.) an exploration of the merits that GWAP can bring to the field of XAI and 2.) a presentation of our approach to design and playtesting of a GWAP for XAI.

## Relevant Work

Explainable AI (XAI) attempts to bring transparency to how AI systems make decisions. The trend toward creating fair and interpretative algorithms [5] arises from decision makers' desire to have AI systems provide reasoning behind the results they generate [3]. In 2018, Abdul et al. presented research which stated that XAI research "tends to neglect the human side of explanations" and questioned "whether they are usable and practical in real-world situations" [5]. This statement is echoed by Zhu et al. who noted that most work in XAI focuses "on new algorithms of XAI rather than on usability, practical interpretability and efficacy on real users" [14].

A recent perspective from social scientists suggests that "most work in explainable artificial intelligence uses only the researchers' intuition of what constitutes a 'good' explanation" [19]. Current evaluation approaches for XAI have been referred to as "you'll know it when you see it" rather than being evidence-based [3].

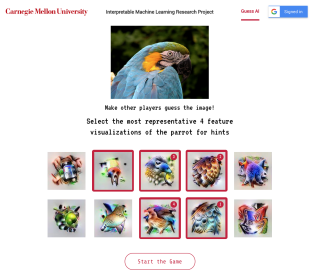


**Figure 1:** The explainer is given a source image (e.g. the parrot). They select the top 4 explanations that they believe will lead other players to guess the correct answer (parrot) as quickly as possible.

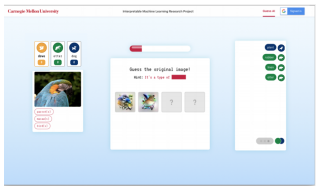

**Figure 2:** Other players who are guessers compete against each other to guess what selected visualizations represent.

Initial evaluations in Human Computer Interaction suggest that explanations have value to users who rely on machine learning [11]. Interactive visual analytic systems offer promise in making complex AI more transparent by providing interaction techniques to reveal insights about decisions [12]. A variety of visual systems have been proposed to help AI algorithms be more clear for users [8, 18, 27, 16]. However, these evaluations typically focus on task-specific tools for a small number of users.

## Applying GWAP Design for XAI

Traditionally, GWAP rely on eliciting high-quality player contributions through intrinsic design methods [7], while weeding out bad data [4]. Expanding research on eliciting high-quality data expands the fields of both GWAP and XAI. Additionally, designing GWAP for XAI is challenging because there is no ground truth to this topic other than player understanding.

*Divergent and Transactive Approaches*

With GWAP design for XAI, we are interested in exploring the trade-offs of divergent and transactive approaches to elicit player contributions. The divergent approach is drawn from creative theory and values uniqueness in player responses; it is better to get many different contributions and have diverse ideas [22]. The transactive approach builds on learning theory; participants build on each others' ideas and deliver fewer contributions with stronger confidence [10]. In our prototype, we produce opportunities for divergence and transactivity by controlling whether players see others' guesses.

*Useful Mechanics*

From a design perspective, GWAP mechanics which involve action, verification, and feedback allow players to focus on engaging problem solving which lends itself to XAI. Action mechanics allow the player to solve the human-computation problem at hand; verification mechanics collect player data which is structured into task-relevant outcomes (e.g. player has provided good or bad data); and feedback mechanics provide responses to players on both in-game behavior and their task outcome [26].

*Success Metrics*

GWAP are varied in their design and innovation for educational games is ongoing [4, 26, 21]. Metrics for evaluating the success of a GWAP include both task metrics, such as how many tasks were completed in a given time frame, and player engagement metrics, such as how often players return or improve in subsequent games [26]. Success on both task and engagement metrics are critical to GWAP as games should provide a good player experience in addition to collecting data.

*Visualization Techniques*

We designed our first GWAP as an XAI assessment game that focuses on deep learning models for image recognition. Feature visualizations [20] and saliency maps [24] are popular visualization techniques used by researchers and practitioners to explain image recognition-tasks. These techniques can yield intuitive and occasionally beautiful visual representations that provide clues that neural networks may be behaving properly, making "the hidden layers comprehensible" [18]. These examples, sometimes selected by model builders to demonstrate efficacy of neural networks, may not be useful in providing evidence that the neural networks perform as expected, as it is difficult to assess the scope and quality of explanations [15]. Therefore, we identified image recognition as a relevant area for designing scalable validation techniques.


**Figure 3:** If the guessers are unable to guess the image after the 4 explanations are revealed, they receive a text hint.

## Designing the Game: XAI for Image Recognition

To demonstrate the value of GWAP for XAI, we designed a multi-player GWAP for image recognition. The game was implemented in JavaScript using the React library [1].

*Learned Feature Visualizations*

Feature visualizations are produced by making neural network interpretation of images visible. Over the course of multiple layers, abstractions are built from detected edges, textures, patterns, and parts of an image. We computed feature visualizations from images of animals and objects using the Lucid library [2].

*Player Roles*

One player, the "explainer," is given a source image (Figure 1). The explainer selects the top explanations that they believe will lead the other players to guess the correct answer (e.g. parrot) as quickly as possible. Two other players are "guessers," who compete against each other to guess what the feature visualizations represent.

*Points as Incentives*

Guessers receive one visual explanation to start, with a new explanation revealed every fifteen seconds (e.g. 2 of the 4 visualizations are revealed to guessers in Figure 2). The quicker a guesser identifies the correct answer, the more points the guesser and explainer gain. If the guessers are unable to guess the image (e.g. parrot) after all four images are revealed, they receive a text hint (Figure 3). We utilize transactive design, where all guesses are visible to the explainer and to the guessers. This allows each guesser to build on their and other players' guesses. Providing accurate data is the ideal way to play this game, because both explainers and guessers aim to get to a correct guess as quickly as possible. In particular, we hypothesize that guessers are highly motivated not only by being quick to guess, but by being the first to guess, as only one of the two guessers receives points. Early pilot testing suggested that players find reaching agreement quickly to be satisfying and valuable.

*Data Collection*

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 our algorithms, and the explainer is only allowed to select four of the ten. Explainers also dictate the order in which the images are shown to guessers. Of the four visualizations they select, they are instructed to select the most helpful explanation first. Second, the guessers type guesses conveying how they interpret the visualizations and give information about which image(s) help them guess correctly. Guessers provide copious and timely data: they can guess as often as they want within a time limit.

*Adjusting Parameters*

Through versions of the game, we can capitalize on opportunities for divergence and transactivity by controlling whether the players can see each others' guesses. If the explainer can see all guesses, but players can only see their own, we test a divergence-supporting system; if players can see each others' guesses and are incentivized to build on each other's contributions, we test a transactive-supporting system.

## Evaluating the Game: Playtesting

*Participants*

There are two primary methods for recruiting GWAP studies. The first relies on crowdsourced labor platforms [9] and the second recruits from groups likely to take an interest

[1] https://reactjs.org/
[2] https://github.com/tensorflow/lucid

in the domain [6]. For our initial tests, we recruited participants who are interested in artificial intelligence.

*Procedures*
Participants use computers to play the game. Each playtest session is designed to take no more than 30 minutes. Logging in with an email address is required to play the game; this serves as the player identification and provides the ability to track progress during a single game or over multiple games. As participants play the game, the web server running the game logs player responses in a password-protected database. Participants are asked if they are interested in a post-game interview. This interview provides qualitative feedback which can capture information to supplement log data (e.g. player motivation, points of confusion, additional thoughts).

*Playtesting Tasks*
As part of the research study, players are provided instructions to complete in their role as the 'explainer.' In the first game, the task is to select the animals category, and within the category to chose the parrot. Selection of the parrot will create a high quantity of data for the parrot and create a consistent experience for users in their first playthrough of the game. In the second game, the task is to chose the same category, animals, but pick another image. In the third game, players can choose a different category, object, and any image.

*Measures and Analyses*
Our analysis of our playtesting is currently in-progress. We plan to leverage metrics to analyze the game to understand the effectiveness of the explanations. This includes determining agreement between the explainers and the guessers to understand if explanations that the explainers select lead to correct interpretations by the guessers.

Patterns of agreement can also provide evidence that the explanations make sense to multiple people. Tracking guessers' responses is useful to determine if explanations are ambiguous and lead to a diverse set of guesses, or if they converge on an interpretation. Measuring engagement can tell if users are willing to repeat playing the game. To collect data at scale, keeping the players engaged is critical for use of GWAP for XAI. Tracking engagement will allow us to analyze why users may improve at interpreting explanations over time, by comparing their performance at different levels of their experience with the game.

## Evaluating the Game: Data Analytics

We are currently performing user playtests while scoping milestones and upcoming features. To generate data through playtesting, quantitative and qualitative methods are used.

*Quantitative Data through Web Analytics*
Through quantitative data collection, the category of image and explanations which the explainer chooses to describe the image are tracked along with the order in which the explanations are selected. As explanations are revealed to guessers, guesses are logged and associated information (e.g. how long it takes guessers to identify an image from explanations) is logged. Quantitative data is collected as players complete and replay games. Quantitative data can help answer the following questions about GWAP for XAI image visualization:

- What explanations do explainers choose from those provided?
- How do explainer-selected visualizations align to those that the AI selects as top-ranked explanations for given images?
- Will guessers be able to interpret top-ranked explanations better than lower-ranked explanations?

- How does accuracy of player interpretation of explanations evolve over multiple games (e.g. guessers guessing correct image after fewer reveals and explainers selecting more 'top-ranked' visualizations)?
- What effect might gamification [25] provide to identifying explanations and improving user satisfaction during gameplay (e.g. tracked through points earned)?

*Qualitative Survey Data*

As part of playtests, users are surveyed about their experience. Following the first and second games, clarifying questions are asked to determine areas of uncertainty and gauge user perspective of their performance. Following playtesting, questions are asked to gather additional feedback (e.g. would there be something to motivate you to play more?; would you share your game results with others, what makes this appealing, why or why not?). Users are also surveyed with questions where they quantify an answer: Strongly Disagree, Disagree, Neutral, Agree, Strongly Agree. These survey questions try to capture how users might perceive and experience the game (e.g. I consider myself a gamer; I felt engaged when selecting visualizations; selecting visualizations was intuitive; replaying the game made me more successful in guessing).

## Discussion

Testing our game will provide metrics which can help us assess if GWAP can provide value for XAI. Through deploying our game we: 1.) produce an initial dataset; and 2.) analyze the dataset to understand the impact of specific image explanations on players' ability to identify original images. Collecting data will demonstrate basic validity of our approach of GWAP for XAI that we can engage players and produce usable data. The dataset can also support assessment of how often players agree on what the explanations mean and if selection corresponds with the ground truth of the labeled data from which the explanation was generated.

## Conclusion

This project will contribute significant empirical and technical knowledge of how to create human-powered systems for assessing explainable AI (XAI). This research will contribute to literature on XAI, image recognition, and GWAP. This work also provides empirical evidence on the validity of GWAP designs for XAI. Validation of insights will contribute to how transactive and divergent multiplayer designs can elicit user input.

## Acknowledgements

This work was a collaboration between the Data Interaction Group and the Oh! Labs within the Human-Computer Interaction Institute at Carnegie Mellon University. We thank individuals who participate in our user research, as well as previous contributors Shivang Gupta and Pranav Dheram who contributed to building our GWAP XAI prototype for image recognition.

## References

[1] Luis Von Ahn and Laura Dabbish. 2008. "Designing games with a purpose". *Commun. ACM* 51, 8 (2008).

[2] Vickie Curtis. 2015. Motivation to participate in an online citizen science game: A study of foldit. *Science Communications* 37, 6 (2015), 723–746.

[3] Doshi-Velez and Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. (2017).

[4] Dion Hoe-Lian Goh Ei Pa Pa Pe-Than and Chei Sian Lee. 2015. A typology of human computation games: an analysis and a review of current games. *Behaviour Information Technology* 34 (2015), 809–824.

[5] Abdul et al. Trends and trajectories for explainable, accountable and intelligible systems. In *Proceedings of*

*the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 582.

[6] Chris Madge et al. 2017. Experiment-Driven Development of a GWAP for Marking Segments in Text.

[7] David Gundry et al. 2018. "Intrinsic elicitation: A model and design approach for games collecting human subject data.". *"International Conference on the Foundations of Digital Games"* (2018).

[8] Matthew Kay et al. 2016a. When (Ish) is My Bus?: User-centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. In *CHI Conference on Human Factors in Computing Systems*.

[9] Marta Sabou et al. 2014. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. (2014).

[10] Miaomiao Wen et al. 2016b. Transactivity as a predictor of future collaborative knowledge integration in team-based learning in online courses. *International Educa- tional Data Mining Society* (2016).

[11] Simone Stumpf et al. 2009. Interacting meaningfully with machine learning systems: Three experiments. *Int. J. Hum.-Comput. Stud.* 67 (2009), 639–662.

[12] Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. 2018. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. (2018).

[13] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. 2017. What do we need to build explainable AI systems for the medical domain? (2017).

[14] Sebastian Risi Rafael Bidarra Jichen Zhu, Antonios Liapis and G Michael Youngblood. 2018. Explainable AI for designers: a human-centered perspective on mixed-initiative co-creation. IEEE, 1–8.

[15] Michael Muelly Ian Goodfellow Moritz Hardt Julius Adebayo, Justin Gilmer and Been Kim. 2018. Sanity checks for saliency maps. (2018).

[16] Josua Krause, Adam Perer, and Kenney Ng. 2016. Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models. In *2016 CHI Conference on Human Factors in Computing Systems*.

[17] Edith L. M. Law, Luis Von Ahn, Roger B. Dannenberg, and Mike Crawford. 2007. Tagatune: A game for music and sound annotation. (2007).

[18] Xu Wang Steven P Dow James Herbsleb Miaomiao Wen, Keith Maki, Ludwig Schubert Ian Johnson Carolyn RoShan Carter, Zan Armstrong, and Chris Olah. 2019. Exploring Neural Networks with Activation Atlases. (2019).

[19] Tim Miller. 2017. Explanation in Artificial Intelligence: Insights from the Social Sciences. (2017).

[20] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. 2017. Feature Visualization. *Distill* (2017).

[21] 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. (2017).

[22] Mark A Runco. 2010. Divergent thinking, creativity, and ideation. *The Cambridge handbook of creativity* 413 (2010). Issue 446.

[23] Nitin Seemakurty, Jonathan Chu, Luis von Ahn, and Anthony Tomasic. 2010. Word Sense Disambiguation via Human Computation. In *ACM SIGKDD Workshop on Human Computation (HCOMP '10)*.

[24] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. (2013).

[25] 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 (CHI PLAY '16)*.

[26] Kristin Siu, Alexander Zook, and Mark O. Riedl. 2017. A Framework for Exploring and Evaluating Mechanics in Human Computation Games. In *International Conference on the Foundations of Digital Games*.

[27] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. 2020. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 56–65.