On Design and Evaluation of Human-centered Explainable AI systems

Upol Ehsan
Georgia Institute of Technology, Atlanta, GA
Cornell University, Ithaca, NY
ehsanu@gatech.edu

Mark O. Riedl
Georgia Institute of Technology, Atlanta, GA
riedl@cc.gatech.edu

ABSTRACT

As AI systems become ubiquitous in our lives, the human side of the equation needs careful investigation. The challenges of designing and evaluating "black-boxed" AI systems depends crucially on *who* the human is in the loop. Explanations, viewed as a form of post-hoc interpretability, can help establish rapport, confidence, and understanding between the AI agent and the user, especially when it comes to understanding failures and unexpected AI behavior. To effectively design and evaluate explanation generation systems, we need deeper end-to-end investigations of incorporating fully-realized AI agents and automated explanation generation systems into user studies. In this paper, we present a case study that focuses on how non-expert users perceive different styles of automatically generated rationales by an AI agent along the dimensions of *confidence*, *humanlike-ness*, *adequate justification*, and *understandability*. We summarize our results and provide a desiderata of research questions yet to be addressed.

CCS CONCEPTS

• **Human-centered computing** → **HCI design and evaluation methods**; • **Computing methodologies**;

KEYWORDS

Explainable AI, rationale generation, user perception, algorithmic decision-making, interpretability, Artificial Intelligence, Machine Learning

Glasgow '19, Glasgow, Scotland,
2019. ACM ISBN xxxx…$xxxx
https://doi.org/xxxx

INTRODUCTION

Explainable AI (XAI) refers to artificial intelligence (AI) and machine learning techniques that can provide human understandable justification for their output behavior. Much of the previous work on explainable AI has focused on *interpretability*. We view interpretability as a property of machine learned models that dictate the degree to which a human user—AI expert or non-expert user—can come to conclusions about the performance of the model on specific inputs. *Explanation generation* can be described as a form of post-hoc interpretability; an important distinction between interpretability and explanation generation is that explanation does not necessarily elucidate precisely how a model works but aims to give useful information for practitioners and users in an accessible manner.

Semi- and fully-autonomous AI systems are becoming commonplace in all aspects of our lives, from voice assistants that we interact with directly, to hiring and predictive policing algorithms that operate without our direct awareness. As the complexity of the AI systems and algorithms grow, we increasingly come to think of them as "black boxes" that defy understanding in the sense that increasing amount of technical expertise and effort are required to discern why an AI system made a decision or performed a behavior. The issues pertaining to AI decision-making being a "black box" are significantly exacerbated when dealing with non-expert users that are increasingly required to interact with AI systems. Explainability is an important as a means to build rapport, confidence, and understanding between the AI agent and its user, especially when it comes to understanding failure and unexpected behavior.

At the heart of explainability is meaning-making. The explain-"ability" of a computing system is often reliant on the human's ability to make sense of its working. Thus, the meaning-making is a relational process where the alignment of situated epistemologies of the user and the machine needs to take place. However, frequently lost in the technical discussion of explainability in AI is the *human* side of the issue, which is where the Human Computer Interaction (HCI) community can play a vital role. Explainability in AI is as much as HCI's problem as it is AI's. The HCI community has started looking at the human-centered design and evaluation challenges of black-boxed systems [1]. However, end-to-end studies incorporating fully-realized AI agents and explanation generation systems into human-subject studies can facilitate better design and evaluation of XAI systems. In this paper, we share human-centered insights from our prior work [2] studying how automatically generated explanations affect user perceptions and evaluation of an XAI system. We share a data collection pipeline that creates a corpus of natural language "think-aloud" explanation data followed by insights from two user studies focusing on the measures of evaluation along dimensions of human factors. Finally, we reflect on the human-centered aspects of the case study and put forth a desiderata of research questions that the HCI community is uniquely positioned to tackle

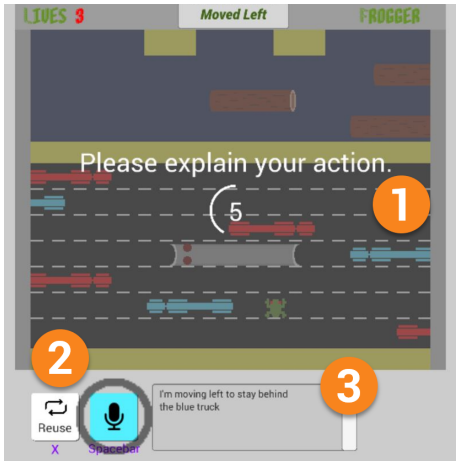

Figure 1: The rationale collection process. (1) Game pauses after each action. (2) Speech-to-text transcribes the rationale. (3) Participants can view and edit the transcribed rationales.

CASE STUDY

One viable approach to post-hoc explanation generation is *automated rationale generation*, a process of producing a natural language explanation for agent behavior *as if a human had performed the behavior* [2]. The intuition behind rationale generation is that humans can engage in effective communication by verbalizing plausible motivations for their action, even when the verbalized reasoning does not have a consciously accessible neural correlate of the decision-making process [3]. Whereas an explanation can be in any communication modality, rationales are natural language explanations. Natural language is arguably the most accessible modality of explanation. However, since rationales are natural language explanations, there is a level of abstraction between the words that are generated and the literal inner workings of an intelligent system. This motivates a range of research questions pertaining to how the choice of words for the generated rationale affect human-factor dimensions such as confidence in the agent's decision, understandability, human-likeness, explanatory power, tolerance to failure and unexpected behavior, likeability, and perceived intelligence.

In this case study, the automated rationale generation is treated as the problem of translating the internal state and action representations into natural language using a deep neural network trained on human explanations. Beyond the technical implementation, the question of where to get the data, how the data is collected, how the data is used by the algorithm, and who the intended recipient of the generated explanation is can influence the measures to evaluate the success of an XAI system. The case study is conducted in the context of explaining the decisions of an AI agent that plays a game of Frogger, a sequential decision-making task where past actions affect future actions; sequential tasks are typically overlooked in explainable AI research.

In the next sections we give a brief overview of our study [2], what we learned about how explanations affect human-factor dimensions, and, more importantly, what we still need to know in order to effectively design human-centered explainable AI systems.

Data Collection

There is no readily available dataset for the task of learning to generate explanations. Thus, we developed a methodology to remotely collect live "think-aloud" data from players as they played through a game. To get a corpus of linked game states, actions, and explanations, we built a modified version of Frogger (a sequential environment) in which players simultaneously play the game and explain each of their actions. The entire process is divided into three phases: (1) A guided tutorial, (2) rationale collection, and (3) transcribed explanation review. The guided tutorial (1) ensures that users are familiar with the interface and its use before they begin providing explanations. For rationale collection (2), participants engage in a turn-taking experience where they take an action and explain it while the game is paused (Figure 1). During thinking out loud, an automatic speech-to-text library

Table 1: Examples of rationales generated for the same game action.

Action	Focused-view	Complete-view
Right	I had cars to the left and in front of me so I needed to move to the right to avoid them.	I moved right to be more centered. This way I have more time to react if a car comes from either side.
Up	The path in front of me was clear so it was safe for me to move forward.	I moved forward making sure that the truck won't hit me so I can move forward one spot.
Left	I move to the left so I can jump onto the next log.	I moved to the left because it looks like the logs and top or not going to reach me in time, and I'm going to jump off if the law goes to the right of the screen.
Down	I had to move back so that I do not fall off.	I jumped off the log because the middle log was not going to come in time. So I need to make sure that the laws are aligned when I jump all three of them.

transcribes the utterances, substantially reducing participant burden and making the flow more natural than having to type down their utterances. Upon game play completion (3), players review all action-explanation pairs in a global context by replaying each action [Figure 1]. While we use Frogger as a test environment in our experiments, a similar user experience can be designed using other turn-based environments. We deployed our data collection pipeline on *Turk Prime* and collected over 2000 samples of action-explanation pairs from 60 participants.

Neural Translation Model

We use an encoder-decoder recurrent neural network [4] to teach our network to generate relevant natural language explanations for any given action. These kinds of networks are commonly used for machine translation tasks or dialogue generation, but their ability to understand sequential dependencies between the input and the output make it suitable for explanation generation in sequential domains. Specifically, the network learns how to translate the input game state representation, comprised of the representation of the game combined with other influencing factors, into an output rationale as a sequence of words. Thus our network learns to translate game state and action information into natural language rationales.

To experiment with different strategies for rationale generation, we vary the input configurations with the intention of producing varying styles of rationales. Empirically, we found that machine learning agents can play the game effectively when given a limited portion of the screen. Thus a natural configuration for the rationale generator is to give it the same observation window that the agent needs to learn to play. This *focused-view* configuration, while accurately reflects what the agent is considering also leads to concise rationales due to the limitation of data it has available for rationale generation Thus we formulated a second *complete-view* configuration that gives the rationale generator the ability to use all information on the board. This produces more detailed, holistic rationales and uses state information the algorithm is not considering. See Table 1 for example rationales generated by our system.

User Study Highlights

We conducted two user studies to evaluate the generated rationales, especially how they affect human perceptions (see [2] for details). In both user studies, participants watched videos (Figure 2) for details where the agent is taking a series of actions and "thinking out loud" in different styles. The first user study establishes the viability of the generated rationales and situates user perception along the dimensions of *confidence*, *human-likeness*, *adequate justification*, and *understandability*. We adapted these constructs from technology acceptance models and related research in HCI [5]. Analyzing the open-ended justifications using a combination of thematic analysis and grounded theory, we found emergent components that speak to each dimension [see [2] for details of the analysis]. For

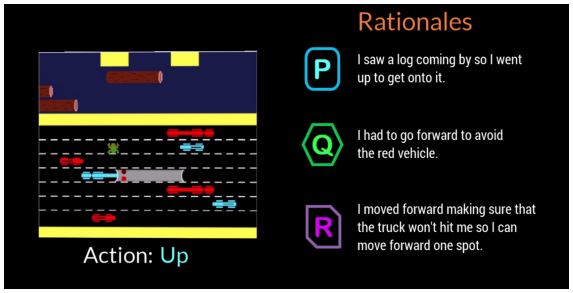

Figure 2: User study screenshot depicting the action and the rationales: *P = Random(lower baseline), Q = Exemplary (higher baseline), R = Candidate*

Table 2: Descriptions for the emergent *components* underlying the human-factor *dimensions* of the generated rationales. [see [2] for details]

Component	Description
Contextual Accuracy	Accurately describes pertinent events in the context of the environment.
Intelligibility	Typically error-free and is coherent in terms of both grammar and sentence structure.
Awareness	Depicts and adequate understanding of the rules of the environment.
Relatability	Expresses the justification of the action in a relatable manner and style.
Strategic Detail	Exhibits strategic thinking, foresight, and planning.

confidence, participants find that contextual accuracy, awareness, and strategic detail are important to have faith in the agent's ability to do its task. Whether the generated rationales appear to be made by a human (*human-likeness*) depend on their intelligibility, relatability, and strategic detail. In terms of explanatory power (*adequate justification*), participants prefer rationales with high levels of contextual accuracy and awareness. For the rationales to convey the agent's motivations and foster *understandability*, they need high levels of contextual accuracy and relatability (see Figure 3 for a mapping and [2] for details).

The second user study builds on the first one and further explores user preferences between the generated rationales with regard to *confidence* in the autonomous agent, communicating *failure* and *unexpected behavior*. We find alignment between the intended differences in features of the generated rationales and the perceived differences by users. Participants were unaware of the correspondence of the network configuration and the style of rationale. Without any knowledge beyond what is shown on the video, they described the difference in the styles of the rationales in a way that was consistent with the intended differences between them. In order to form a stable mental model of the agent's behavior, we find that the level of detail in explanations plays a nuanced role especially when communicating failure and unexpected behavior.

REFLECTIONS

At first glance, it may appear that Frogger is a non-representative real-world domain for exploring automated rationale generation. However, therein lies the point—considering issues of fairness, accountability, and transparency of sociotechnical systems, it is risky to directly test out these systems in mission critical domains without a formative and substantive understanding of the human factors around XAI systems. By conducting the case study in a controlled setting as a first step, we obtain a formative understanding of the technical and the human side, which can then be utilized to better implement such systems in the wild. Subsequent empirical and theoretical work can then build on the transferable insights from this work. Moreover, we can use the human-centered insights to design and evaluate real-world sequential decision-making tasks (e.g., cybersecurity threat validation), where a sequence of non-trivial actions leads to a goal, similar to Frogger's environment.

The case study also catalyzes future exploration around the *who* the human is in AI-powered system design. First, the *who* governs *how* the data is collected, *what* data can be collected, and the most effective way describing the *why* behind an action. The communities of practice interfacing with the systems circumscribe the domain of explanations requisite for an explainable system. For instance, if oncologists are using AI-aided decision support systems for diagnosis, they will need rationales that peer oncologists would provide. However, if the same decision is to be relayed to a lay-patient (e.g., an app that explains test results), then the explanation needs to be in an accessible manner. Thus, we will need the right *contextual* data in the corpus to effective produce viable and satisfying rationales.

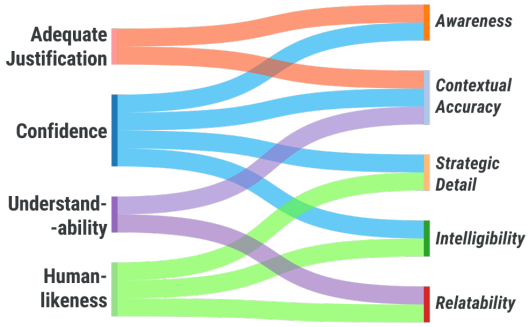

Figure 3: Emergent relationship between the dimensions (left) and components (right) of user perceptions and preference

Focusing on the *who* is a crucial part of the process, where a mixture of qualitative (e.g., ethnographic) and quantitative methods will be needed to devise better data collection and evaluation methods.

Second, the *who* also fundamentally influences the relative importance of the human-centered qualities of an explainable system and its evaluation metrics. Depending on the context, we may need to produce agents of a certain persona; for instance, a companion agent that is concise in communication given time-sensitive nature of operations. However, optimization comes with costs. For instance, conciseness can improve intelligibility but may led to loss in strategic detail, which can hurt confidence in the agent.

Building on these reflections, we briefly outline research questions that have come to light:

- *Timing and frequency:* should we explain every action or only when asked? When and how often are explanations necessary?
- *Temporal evolution:* how should explanations evolve over time? That is, how does long-term explainable systems look like?
- *Interaction paradigm:* instead of the 1-1 interaction, what happens when teams of people interact with rationale generating systems? How does collaborative XAI systems look like?

Explanations—post-hoc interpretability—will play an instrumental role in making systems accessible as AI continues to proliferate complex and sensitive sociotechnical systems. Systematic human-subject studies are necessary to understand how different strategies for explanation generation affect end-users, especially non-AI-experts. End-to-end studies with fully realized XAI systems are needed to determine whether we can control explanation generation to produce the desired down-stream effects on the intended recipients. It is important that we allow ourselves time to work on these issues because the speed of algorithmic development will likely be higher than the rate of societal adaptation. HCI in AI can and should play a calibrating and moderating role in this journey to mitigate unintended consequences and facilitate inclusion of diverse voices in the design of the future of AI.

REFERENCES

[1] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 377.

[2] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark Riedl. 2019. Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions. In *Proceedings of the International Conference on Intelligence User Interfaces*.

[3] Jerry A Fodor. 1994. *The elm and the expert: Mentalese and its semantics*. MIT press.

[4] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025* (2015).

[5] Viswanath Venkatesh, Michael G Morris, Gordon B Davis, and Fred D Davis. 2003. User acceptance of information technology: Toward a unified view. *MIS quarterly* (2003), 425–478.