2018

Learning to Generate Natural Language Rationales for Game Playing Agents

Upol Ehsan
Cornell University

Pradyumna Tambwekar
Georgia Institute of Technology

Larry Chan
Georgia Institute of Technology

Brent Harrison
University of Kentucky, brent.harrison@uky.edu

Mark O. Riedl

Learning to Generate Natural Language Rationales for Game Playing Agents

Learning to Generate Natural Language Rationales for Game Playing Agents

Upol Ehsan[*‡‡], **Pradyumna Tambwekar**[*†], **Larry Chan**[†],
Brent Harrison[‡], and **Mark O. Riedl**[†]
[‡‡]Department of Information Science, Cornell University
[†]School of Interactive Computing, Georgia Institute of Technology
[‡]Department of Computer Science, University of Kentucky

Abstract

Many computer games feature non-player character (NPC) teammates and companions; however, playing with or against NPCs can be frustrating when they perform unexpectedly. These frustrations can be avoided if the NPC has the ability to explain its actions and motivations. When NPC behavior is controlled by a black box AI system it can be hard to generate the necessary explanations. In this paper, we present a system that generates human-like, natural language explanations—called *rationales*—of an agent's actions in a game environment regardless of how the decisions are made by a black box AI. We outline a robust data collection and neural network training pipeline that can be used to gather think-aloud data and train a rationale generation model for any similar sequential turn based decision making task. A human-subject study shows that our technique produces believable rationales for an agent playing the game, *Frogger*. We conclude with insights about how people perceive automatically generated rationales.

Introduction

Non-player characters (NPCs) are interactive, autonomous agents that play critical roles in most modern video games, and are often seen as one crucial component of an engaging player experience. As NPCs are given more autonomy to make decisions, the likelihood that they perform in an unexpected manner increases. These situations risk interrupting a player's engagement in the game world as they attempt to justify the reasoning behind the unexpected NPC behavior. One method to address this side-effect of increased autonomy is to construct NPCs that have the ability to explain their own actions and motivations for acting.

The generation of natural language explanations for autonomous agents is challenging when the agent is a black-box AI, meaning that one doesn't have access to the agent's decision-making process. Even if access were possible, the mapping between inputs and decisions could be difficult for people to interpret. Work by Ehsan *et al.* [8] showed that machine learning models can be trained to provide relevant and satisfactory rationales for their actions using examples of human behavior and human-provided explanations. This is a potentially powerful tool that could be used to create NPCs that can provide human understandable explanations for their own actions, without changing the underlying decision-making algorithms. This in turn could give users more confidence in NPCs and game playing agents and make NPCs and agents more understandable and relatable.

In the work by Ehsan *et al.*, however, the rationale generation model was trained using a semi-synthetic dataset by developing a grammar that could generate variations of actual human explanations to train their machine. While their results were promising, creating the grammar necessary to construct the requisite training examples is a costly endeavor in terms of authorial effort. We build on this work by developing a pipeline to *automatically* acquire a corpus of human explanations that can be used to train a rationale generation model to explain the actions of NPCs and game playing agents. In this paper, we describe our automated explanation corpus collection technique, neural rationale generation model, and present the results of a human-subjects study of human perceptions of generated rationales in the game, *Frogger*.

Related Work

Adaptive team-mate/adversary cooperation in games has often been explored through the lens of decision making [2]. Researchers have looked to incorporate adaptive difficulty in games (cf. [3, 16]) as well as build NPCs which evolve by learning a player's profile as ways to improve the players experience [7, 15]. What is missing from this analysis is the conversational engagement that comes with collaborating with another human player.

NPCs that can communicate in natural language have previously been explored using classical machine learning techniques. These methods often undertake a rule based or probabilistic modeling approach. Buede et al. combine natural language processing with dynamic probabilistic models to maximize rapport between two conversing agents [6]. Prior work has also shown the capacity to use a rule-based system to create a conversational character generator [12]. Both of these methods, however, have a high degree of hand-authoring involved in generating these models. Our work can generate NPCs with similar communicative capabilities with minimal hand-authoring.

[*]Denotes equal contribution.

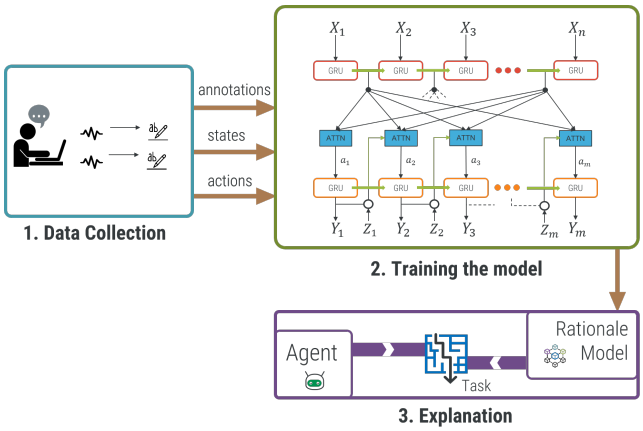


Figure 1: End to End Pipeline for training a system that can generate explanations.

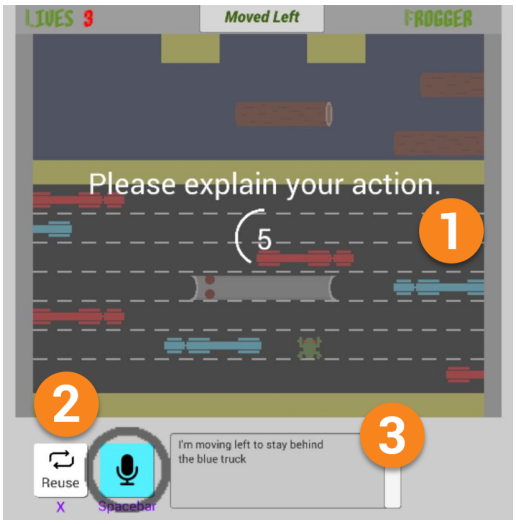


Figure 2: Players take an action and verbalize their rationale for that action. (1) After taking each action, the game pauses for 10 seconds. (2) Speech-to-text transcribes the participant's rationale for the action. (3) Participants can view their transcribed rationales near-real time and edit them, if needed.

Explainable AI has attracted interest from researchers across various domains. The authors of [1] conduct a comprehensive survey on burgeoning trends in explainable and intelligible systems research. Certain intelligible systems researchers look to use model-agnostic methods to add transparency to the latent technology [13, 17]. Other researchers use visual representations to interpret the decision-making process of a machine learning system [9]. We situate our system as an agent that unpacks the thought process of a human player, if they were to play the game.

Evaluation of explainable AI systems can be difficult because the appropriateness of an explanation is subjective. One approach to evaluating such systems was proposed in [5]. They presented participants with different fictionalized explanations for the same decision and measured perceived levels of justice among their participants. We adopt a similar procedure to measure the quality of generated rationales versus alternate baseline rationales.

Learning to Generate Rationales

We define a *rationale* as an explanation that justifies an action based on how a human would think. These rationales do not reveal the true decision making process of an agent, but still provide insights about why an agent made a decision in a form that is easy for non-experts to understand.

Rationale generation requires translating events in the game environment into natural language outputs. Our approach to rationale generation involves two steps: (1) collect a corpus of think-aloud data from players who explained their actions in a game environment; and (2) use this corpus to train an encoder-decoder network to generate plausible rationales for any action taken by an agent (see Figure 1).

Data Collection Interface

There is no readily available dataset for the task of learning to generate explanations. Thus, we developed a methodology to collect live "think-aloud" data from players as they played through a game. This section covers the two objectives of our data collection endeavor:

1. Create a think-aloud protocol in which players provide natural rationales for their actions.
2. Design an intuitive player experience that facilitates accurate matching of the participants' utterances to the appropriate state in the environment.

To train an agent to generate rationales we need data linking game states and actions to their corresponding natural language explanations. To achieve this goal, we built a modified version of Frogger in which players simultaneously play the game and also 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.

During the guided tutorial, our interface provides instruction on how to play through the game, how to provide natural language explanations, and how to review/modify any explanations they have given. This helps ensure that users are familiar with the interface and its use before they begin providing explanations.

During explanation collection, users play through the game while explaining their actions out loud. Figure 2 shows the game embedded into the explanation collection interface. To help couple explanations with actions, the game pauses for 10 seconds after an action is taken. During this time, the player's microphone automatically turns on and the player is asked to explain their most recent action while a speech-to-text library transcribes the explanation.

Participants can view their transcribed text and edit it if necessary. During preliminary testing, we observed that players often repeat a move and the explanation is the same. For ease, participants can indicate that the explanation

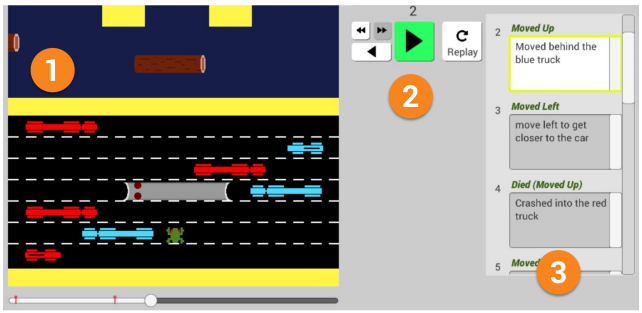


Figure 3: Players can step-through each of their action-rationale pairs and edit if necessary. (1) Players can watch a replay of their actions while editing their rationales. (2) Players use these buttons to control the flow of their step-through. (3) The rationale for the current action gets highlighted for review.

accompanying their most recent explanation is the same as that of the last action performed.

During transcribed explanation review, users are given one final opportunity to review and edit the explanations given during gameplay (see Figure 3). Players can step through all of the actions they performed in the game and see their accompanying transcribed explanations so they can see the game context in which their explanations were given.

The interface is designed so that no manual hand-authoring/editing of our data was required before pushing it into our machine learning model. Throughout the game, players were given the opportunity to organically edit their own data without impeding their work-flow. This added layer of frictionless editing was crucial in ensuring that we can directly input the collected data into the network with zero manual cleaning.

One core strength that facilitates transferability is that our pipeline is environment and domain agnostic. While we use Frogger as a test environment in our experiments, a similar user experience can be designed using other turn-based environments with minimal effort.

Neural Translation Model

We use an encoder-decoder network 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 out task. Our encoder decoder architecture is similar to that used in [8]. The network learns how to translate the input game state representation $X = x_1, x_2, ..., x_n$, comprised of the sprite representation of the game combined with other influencing factors, into an output explanation as a sequence of words $Y = y_1, y_2, ..., y_m$ where y_i is a word. The input X has a fixed size of 261 tokens encompassing the game state representation, lives left and the location of the frog. The vocabulary sizes for the encoder and the decoder are 491 and 1104 respectively. Thus our network learns to translate game state and action information into natural language rationales.

The encoder and decoder are both recurrent neural networks (RNN) comprised of GRU cells. The decoder network uses an additional attention mechanism [11] to learn to weight the importance of different components of the input with regard to their effect on the output.

To simplify the learning process, the state of the game environment is converted into a sequence of symbols where each symbol represents a type of sprite. To this, we append information concerning Frogger's position, the most recent action taken, and the number of lives the player has left to create the input representation X. On top of this network structure, we vary the input configurations with the intention of producing varying styles of rationales. These two configurations are titled the *focused view* configuration and the *complete-view* configuration and are used throughout the experiments presented in this paper.

Focused-view Configuration In this configuration we used a windowed representation of the grid, i.e. only a 7×7 window around the Frog was used in the input. Both playing an optimal game of Frogger and generating relevant explanations based on the current action taken typically only requires this much local context. Therefore providing the agent with only the window around Frogger helps the agent produce explanations grounded in it's neighborhood. In this configuration we prioritized rationales focused on short term awareness over long term planning.

Complete-view Configuration The *complete-view* configuration is an alternate setup that provides the entire game board as context for the rationale generation. There are two differences between this configuration and the focused-view configuration. First, instead of showing the network only a window of the game, we use the entire game screen as a part of the input. The agent now has the opportunity to learn which other long-term factors in the game may influence it's rationale. Second, we added noise to each game state to force the network to generalize when learning to generate rationales and give the model equal opportunity to consider factors from all sectors of the game screen. In this case noise was introduced by replacing input grid values with dummy values. For each grid element, there was a 20% chance that it would get replaced with a dummy value.

Human Perception of Rationales Study

In this section, we attempt to assess whether the rationales generated by our system outperform baselines. We further attempt to understand the underlying components that influence the difference in the perceptions of the generated rationales along four dimensions of human factors: *confidence*, *human-likeness*, *adequate justification*, and *understandability*. Frogger is a good candidate for our experimental design of a rationale generation pipeline for general sequential decision making tasks because it is a simple Markovian environment; that is, the reasons for each action can be easily separated, making it an ideal stepping stone towards a real world environment.

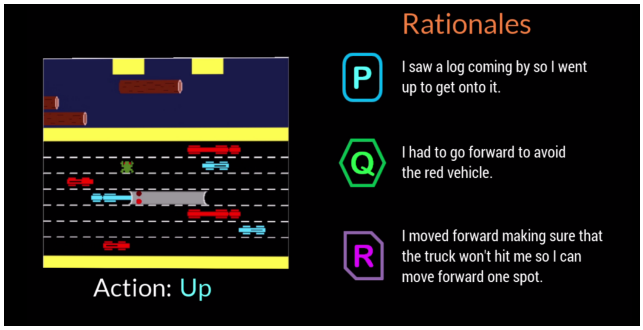


Figure 4: Screenshot from user study (setup 2) depicting the action taken and the rationales: *P = Random, Q = Exemplary, R = Candidate*

To gather the training set of game state annotations we deployed our data collection pipeline on Amazon Turk Prime [10]. From 60 participants we collected over 2000 samples of human explanations corresponding to images of the game when the explanations were made. This comprised the corpus with which we trained the encoder-decoder rationale generation network. The parallel corpus of the collected game state images and natural language rationales was used to train the encoder-decoder network. Each RNN in the encoder and the decoder was parameterized with GRU cells with a hidden vector size of 256. The entire encoder-decoder network was trained for 100 epochs.

We recruited an additional 128 participants, split into two experimental groups of our study through *TurkPrime* [10]; Group1 (age range = 23 - 68, M = 37.4, SD = 9.92), Group 2 (age range = 24 - 59, M = 35.8, 7.67). Forty six percent of our participants were women and only two countries, United States and India, were reported when participants were asked which country they were from. 93% percent of all 128 participants reported that they resided in the United States.

Procedure

Participants watched a series of five videos, each containing an action taken by an agent playing Frogger. In each video, the action was accompanied by three rationales generated by three different techniques (see Figure 5):

- The *exemplary rationale* is the rationale from our corpus that 3 researchers unanimously agreed on as the best one for a particular action. Researchers independently selected rationales they deemed best and iterated till consensus was reached.
- The *candidate rationale* is the rationale produced by our network, either the focused-view or complete-view configuration. This is provided as an upper-bound for contrast with the next two techniques.
- The *random rationale* is a randomly chosen rationale from our corpus.

For each rationale, participants used a 5-point Likert scale to rate their endorsement of each of following four statements, which correspond to four dimensions of interest.

D1. *Confidence:* This rationale makes me confident in the character's ability to perform it's task.

D2. *Human-likeness:* This rationale looks like it was made by a human.

D3. *Adequate justification:* This rationale adequately justifies the action taken.

D4. *Understandability:* This rationale helped me understand why the agent behaved as it did.

Response options on the Likert scale ranged from "strongly disagree" to "strongly agree." In a free-text field, they explained why the ratings they gave for a particular a set of three rationales were similar or different. After answering these questions, they provided demographic information.

Quantitative Results and Analysis

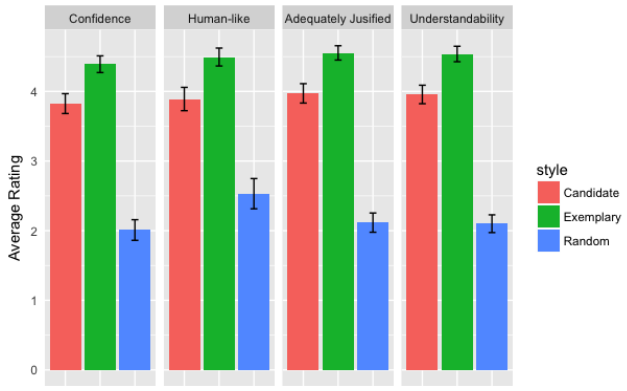
We used a multi-level model to analyze both between-subjects and within-subjects variables. There were significant main effects of rationale style ($\chi^2(2) = 594.80, p < .001$) and dimension ($\chi^2(2) = 66.86, p < .001$) on the ratings. The main effect of experimental group was not significant ($\chi^2(1) = 0.070, p = 0.79$). Figure 5 shows the average responses to each question for the two different experimental groups. Our results support our hypothesis that rationales generated with the focused-view generator and the complete-view generator were judged significantly better across all dimensions than the random baseline ($b = 1.90, t(252) = 8.09, p < .001$). Our results also show that rationales generated by the candidate techniques were judged significantly lower than the exemplary rationale.

The difference between the focused-view candidate rationales and exemplary rationales were significantly greater than the difference between complete-view candidate rationales and exemplary rationales ($p = .005$). Surprisingly, this was because the exemplary rationales were rated lower in the presence of complete-view candidate rationales ($t(1530) = -32.12, p < .001$). Since three rationales were presented simultaneously in each video, it is likely that participants were rating the rationales relative to each other. We also observe that the complete-view candidate rationales received overall higher ratings than the focused-view candidate rationales ($t(1530) = 8.33, p < .001$).

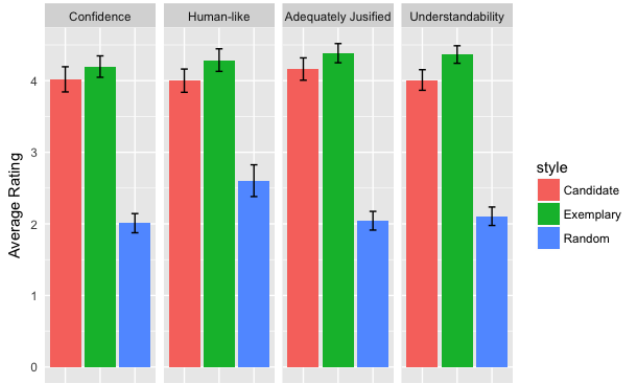
In summary, we established that both the focused-view and complete-view configurations produce believable rationales that perform significantly better than the *random* baseline along four human factors dimensions. While the complete-view candidate rationales were judged to be preferable overall to focused-view candidate rationales, we did not compare them to directly to each other because stylistically one technique may be better suited based on the task and/or game. Our between-subjects study methodology are suggestive but cannot be used to prove any claims between the two experimental conditions.

Qualitative Analysis

In this section, we look at the open-ended responses provided by our participants to better understand the



(a) Focus-View condition.



(b) Complete-View condition.

Figure 5: Human judgment results.

criteria that participants used when making judgments about the *confidence, human-likeness, adequate justification,* and *understandability* of generated rationales. These situated insights augment our understanding of rationale generating systems, enabling us to design better ones in the future.

We analyzed the open-ended justifications participants provided using a combination of thematic analysis [4] and grounded theory [14]. We developed codes that addressed different types of reasonings behind the ratings of the four dimensions under investigation. Next, the research team clustered the codes under emergent themes, which form the underlying *components* of the dimensions. Iterating until consensus was reached, researchers settled on the most relevant five components: (1) *Contextual Accuracy*, (2) *Intelligibility*, (3) *Awareness*, (4) *Relatability*, and (5) *Strategic Detail* (see Table 1). At varying degrees, multiple components influence more than one dimension; that is, there isn't a mutually exclusive one-to-one relationship between components and dimensions.

The remainder of this section will share our conclusions about how these components influence the dimensions of the human factors under investigation. When providing examples of our participants' responses, we will refer to them using the following notation; P1 corresponds to participant 1, P2 corresponds to participant 2, etc.

Table 1: Descriptions for the emergent *components* underlying the human-factor *dimensions* of the generated rationales.

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 (D1) This dimension gauges the participant's faith in the agent's ability to successfully complete it's task and has *contextual accuracy*, *awareness*, *strategic detail*, and *intelligibility* as relevant components. With respect to *contextual accuracy*, rationales that displayed "...recognition of the environmental conditions and [adaptation] to the conditions" (P22) were a positive influence on confidence ratings, while redundant information such as "just stating the obvious" (P42) hindered confidence ratings.

Rationales that showed *awareness* "...of upcoming dangers and what the best moves to make ...[and] a good way to plan" (P17) inspired confidence from the participants. In terms of *strategic detail*, rationales that showed "...long-term planning and ability to analyze information" (P28) yielded higher confidence ratings compared to those that were "...short-sighted and unable to think ahead" (P14) led to lower perceptions of confidence.

Intelligibility alone, without *awareness* or *strategic detail*, was not enough to yield high confidence in rationales. However, rationales that were not *intelligible* (unintelligible) or coherent had a negative impact on participants' confidence:

> The [random and focused-view rationales] include major mischaracterizations of the environment by referring to an object not present or wrong time sequence, so I had very low confidence. (P66)

Human-likeness (D2) *Intelligibility, relatability,* and *strategic detail* are components that influenced participants' perception of the extent to which the rationales were made by a human. Notably, *intelligibility* had mixed influences on the human-likeness of the rationales depending on what participants thought "being human" entailed. Some perceived humans to be fallible and rated rationales with errors more *humanlike* because rationales "...with typos or spelling errors ...seem even more likely to have been generated by a human" (P19). Conversely, some thought error-free rationales must come from a human, citing that a "computer just does not have the knowledge to understand what is going on" (P24).

With respect to *relatability*, rationales were often perceived as more human-like when participants felt that "it mirrored [their] thoughts" (P49), and "... [layed] things out in a way that [they] would have" (P58). Affective rationales had high *relatability* because they "express human emotions including hope and doubt" (P11).

Strategic planning had a mixed impact on human-likeness just like *intelligibility* as it also depended on participants' perception of critical thinking and logical planning. Some participants associated "... critical thinking [and ability to] predict future situations" (P6) with human-likeness whereas others associated logical planning with non-human-like, but computer-like rigid and algorithmic thinking process flow.

Adequate Justification (D3) This dimension unpacks the extent to which participants think the rationale adequately justifies the action taken and is influenced by *contextual accuracy*, and *awareness*. Participants downgraded rationales containing low levels of *contextual accuracy* such as irrelevant details. As P11 puts it:

> The [random and exemplary rationales] don't pertain to this situation. [The Complete View] does, and is clearly the best justification for the action that Frogger took because it moves him towards his end goal.

Beyond *contextual accuracy*, rationales that showcase *awareness* of surroundings rate high on the *adequate justification* dimension. For instance, P11 rated the *random* rationale low because it showed "no awareness of the surroundings". For the same action, P11 gave high ratings for the *exemplary* and *focused-view* rationales because each made the participant "... believe in the character's ability to judge their surroundings."

Understandability (D4) For this dimension, components such as *contextual accuracy* and *relatability* influence participants' perceptions of how much the rationales helped them understand the motivation behind the agent's actions. *Contextually accurate* rationales were found to have a high influence with understandability. In fact, many expressed how the contextual accuracy, not the length of the rationale, mattered when it came to understandability. While comparing the *exemplary* and *focused-view* rationales for understandability, P41 made a notable observation:

> The [exemplary and focused-view rationale] both described the activities/objects in the immediate vicinity of the frog. However, [exemplary] was not as strong as [focused-view] given the frog did not have to move just because of the car in front of him. [Focused-view] does a better job of providing understanding of the action

Participants put themselves in the agent's shoes and evaluated the understandability of the rationales based on how *relatable* they were. In essence, some asked "Are these the same reasons I would [give] for this action?" (P43). The more relatable the rationale was, the higher it scored for understandability.

Design Implications

The understanding of the *components* and *dimensions* can help us design better autonomous agents from a human factors perspective. These insights can also enable tweaking of the network configuration and reverse-engineering it to maximize the likelihood of producing *rationale sytles* that meet the needs of the task, game, or agent persona.

For instance, given the nature of the inputs, choosing a network configuration similar to the *focused-view* can afford the generation of *contextually accurate* rationales. On the other hand, the *complete-view* network configuration can produce rationales with a higher degree of *strategic detail* that can be beneficial in contexts where detail is important, such an explainable oracle. Moreover, an in-game tutorial or a companion agent can be designed using a network configuration that generates *relatabile* outputs to keep the player entertained and engaged.

Future Work

We can extend our current work in other domains of Explainable AI, exploring applications for other sequential decision making tasks. We also plan to deploy our rationale generator with an collaborative NPC in an interactive game to investigate how the perception of a collaborative agent changes when players interact longitudinally (over an extended period of time). This longitudinal approach can help us understand novelty effects of rationale generating agents. Besides NPCs, our techniaues can improve teaching and collaboration in games, especially around improvisation and co-creative collaboration in game-level designs

Our data collection pipeline is currently designed to work with discrete-action games that have natural break points where the player can be asked for explanations, making it less disruptive than continuous-time and -action games. The next challenge is to extend and test our approach with more continuous spaces where states aren't as well defined and rationales are harder to capture from moment-to-moment.

Conclusions

In this paper, we explore how human justifications for their actions in a video game can be used to train a system to generate explanations for the actions of autonomous game-playing agents. We introduce a pipeline for automatically gathering a parallel corpus of game states annotated with human explanations and show how this corpus can be used to train encoder-decoder networks. The resultant model thus translates the state of the game and the action performed by the agent into natural language, which we call a *rationale*. The rationales generated by our technique are judged better than those of a random baseline and close to matching the upper bound of human rationales. By enabling autonomous agents to communicate about the motivations for their actions, we hope to provide users with greater confidence in the agents while increasing perceptions of understanding and relatability.

References

[1] Ashraf Abdul et al. “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*. ACM. 2018, p. 582.

[2] Aswin Thomas Abraham and Kevin McGee. “AI for dynamic team-mate adaptation in games”. In: *Computational Intelligence and Games (CIG), 2010 IEEE Symposium on*. IEEE. 2010, pp. 419–426.

[3] Maria-Virginia Aponte, Guillaume Levieux, and Stéphane Natkin. “Scaling the level of difficulty in single player video games”. In: *International Conference on Entertainment Computing*. Springer. 2009, pp. 24–35.

[4] J Aronson. *A pragmatic view of thematic analysis: the qualitative report, 2,(1) Spring*. 1994.

[5] Reuben Binns et al. “’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. 2018, p. 377.

[6] Dennis M. Buede, Paul J. Sticha, and Elise T. Axelrad. “Conversational Non-Player Characters for Virtual Training”. In: *Social, Cultural, and Behavioral Modeling*. Ed. by Kevin S. Xu et al. Cham: Springer International Publishing, 2016, pp. 389–399. ISBN: 978-3-319-39931-7.

[7] Silvia Coradeschi and Lars Karlsson. “A role-based decision-mechanism for teams of reactive and coordinating agents”. In: *Robot Soccer World Cup*. Springer. 1997, pp. 112–122.

[8] Upol Ehsan et al. “Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations”. In: *Proceedings of the AAAI Conference on Artificial Intelligence, Ethics, and Society*. Feb. 2018.

[9] Josua Krause, Adam Perer, and Kenney Ng. “Interacting with predictions: Visual inspection of black-box machine learning models”. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM. 2016, pp. 5686–5697.

[10] Leib Litman, Jonathan Robinson, and Tzvi Abberbock. “TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences”. In: *Behavior research methods* 49.2 (2017), pp. 433–442.

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

[12] Grant Pickett, Foaad Khosmood, and Allan Fowler. “Automated generation of conversational non player characters”. In: *Eleventh Artificial Intelligence and Interactive Digital Entertainment Conference*. Vol. 362. 2015.

[13] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “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*. ACM. 2016, pp. 1135–1144.

[14] Anselm Strauss and Juliet Corbin. “Grounded theory methodology”. In: *Handbook of qualitative research* 17 (1994), pp. 273–85.

[15] Chek Tien Tan and Ho-lun Cheng. “Personality-based Adaptation for Teamwork in Game Agents.” In: *AIIDE*. 2007, pp. 37–42.

[16] Sang-Won Um, Tae-Yong Kim, and Jong-Soo Choi. “Dynamic difficulty controlling game system”. In: *IEEE Transactions on Consumer Electronics* 53.2 (2007).

[17] Jason Yosinski et al. “Understanding neural networks through deep visualization”. In: *arXiv preprint arXiv:1506.06579* (2015).