Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions

Upol Ehsan
School of Interactive Computing,
Georgia Institute of Technology
Atlanta, Georgia
Department of Information Science,
Cornell University
Ithaca, New York
ehsanu@gatech.edu

Pradyumna Tambwekar
School of Interactive Computing,
Georgia Institute of Technology
Atlanta, Georgia
ptambwekar3@gatech.edu

Larry Chan
School of Interactive Computing,
Georgia Institute of Technology
Atlanta, Georgia
larrychan@gatech.edu

Brent Harrison
Department of Computer Science,
University of Kentucky
Lexington, Kentucky
harrison@cs.kyu.edu

Mark O. Riedl
School of Interactive Computing,
Georgia Institute of Technology
Atlanta, Georgia
riedl@cc.gatech.edu

ABSTRACT

Automated rationale generation is an approach for real-time explanation generation whereby a computational model learns to translate an autonomous agent's internal state and action data representations into natural language. Training on human explanation data can enable agents to learn to generate human-like explanations for their behavior. In this paper, using the context of an agent that plays *Frogger*, we describe (a) how to collect a corpus of explanations, (b) how to train a neural rationale generator to produce different styles of rationales, and (c) how people perceive these rationales. We conducted two user studies. The first study establishes the plausibility of each type of generated rationale and situates their user perceptions along the dimensions of *confidence*, *humanlike-ness*, *adequate justification*, and *understandability*. The second study further explores user preferences between the generated rationales with regard to *confidence* in the autonomous agent, communicating *failure* and *unexpected behavior*. Overall, we find alignment between the intended differences in features of the generated rationales and the perceived differences by users. Moreover, context permitting, participants preferred detailed rationales to form a stable mental model of the agent's behavior.

CCS CONCEPTS

• **Human-centered computing** → **HCI design and evaluation methods**; • **Computing methodologies** → *Natural language generation.*

KEYWORDS

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

ACM Reference Format:
Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. 2019. Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions. In *24th International Conference on Intelligent User Interfaces (IUI '19), March 17–20, 2019, Marina del Rey, CA, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3301275.3302316

1 INTRODUCTION

Explainable AI refers to artificial intelligence and machine learning techniques that can provide human understandable justification for their behavior. Explainability is important in situations where human operators work alongside autonomous and semi-autonomous systems because it can help build rapport, confidence, and understanding between the agent and its operator. In the event that an autonomous system fails to complete a task or completes it in an unexpected way, explanations help the human collaborator understand the circumstances that led to the behavior, which also allows the operator to make an informed decision on how to address the behavior.

Prior work on explainable AI (XAI) has primarily focused on non-sequential problems such as image classification and captioning [40, 42, 44]. Since these environments are episodic in nature, the model's output depends only on its input. In sequential environments, decisions that the agent has made in the past influence future decisions. To simplify this, agents often make locally optimal decisions by selecting actions that maximize some discrete notion of expected future reward or utility. To generate plausible explanations in these environments, the model must unpack this local reward or utility to reason about how current actions affect future actions. On top of that, it needs to communicate the reasoning in a human understandable way, which is a difficult task. To address this

IUI '19, March 17–20, 2019, Marina del Rey, CA, USA

ACM ISBN 978-1-4503-6272-6/19/03…$15.00
https://doi.org/10.1145/3301275.3302316

challenge of human understandable explanation in sequential environments, we introduce the alternative task of rationale generation in sequential environments.

Automated rationale generation is a process of producing a natural language explanation for agent behavior *as if a human had performed the behavior* [17]. The intuition behind rationale generation is that humans can engage in effective communication by verbalizing plausible motivations for their action. The communication can be effective even when the verbalized reasoning does not have a consciously accessible neural correlate of the decision-making process [8, 9, 19]. Whereas an explanation can be in any communication modality, rationales are natural language explanations that don't literally expose the inner workings of an intelligent system. Explanations can be made by exposing the inner representations and data of a system, though this type of explanation may not be accessible or understandable to non-experts. In contrast, contextually appropriate natural language rationales are accessible and intuitive to non-experts, facilitating understanding and communicative effectiveness. Human-like communication can also afford human factors advantages such as higher degrees of satisfaction, confidence, rapport, and willingness to use autonomous systems. Finally, rationale generation is fast, sacrificing an accurate view of agent decision-making for real-time response, making it appropriate for real-time human-agent collaboration. Should deeper, more grounded and technical explanations be necessary, rationale generation may need to be supplemented by other explanation or visualization techniques.

In preliminary work [17] we showed that recurrent neural networks can be used to translate internal state and action representations into natural language. That study, however, relied on synthetic natural language data for training. In this work, we explore if human-like plausible rationales can be generated using a non-synthetic, natural language corpus of human-produced explanations. To create this corpus, we developed a methodology for conducting remote think-aloud protocols [20]. Using this corpus, we then use a neural network based on [17] to translate an agent's state and action information into natural language rationales, and show how variations in model inputs can produce two different types of rationales. Two user studies help us understand the perceived quality of the generated rationales along dimensions of human factors. The first study indicates that our rationale generation technique produces plausible and high-quality rationales and explains the differences in user perceptions. In addition to understanding user preferences, the second study demonstrates how the intended design behind the rationale types aligns with their user perceptions.

The philosophical and linguistic discourse around the notion of explanations [26, 31] is beyond the scope of this paper. To avoid confusion, we use the word "rationale" to refer to natural language-based post-hoc explanations that are meant to sound like what a human would say in the same situation. We opt for "rationale generation" instead of "rationalization" to signal that the agency lies with the receiver and interpreter (human being) instead of the producer (agent). Moreover, the word rationalization may carry a connotation of making excuses [30] for an (often controversial) action, which is another reason why we opt for *rationale generation* as a term of choice.

In this paper, we make the following contributions in this paper:

- We present a methodology for collecting high-quality human explanation data based on remote think-aloud protocols.
- We show how this data can be used to configure neural translation models to produce two types of human-like rationales: (1) concise, localized and (2) detailed, holistic rationales. We demonstrate the alignment between the intended design of rationale types and the actual perceived differences between them.
- We quantify the perceived quality of the rationales and preferences between them, and we use qualitative data to explain these perceptions and preferences.

2 RELATED WORK

Much of the previous work on explainable AI has focused on *interpretability*. While there is no one definition of interpretability with respect to machine learning models, we view interpretability as a property of machine learned models that dictate the degree to which a human user—AI expert or user—can come to conclusions about the performance of the model on specific inputs. Some types of models are inherently interpretable, meaning they require relatively little effort to understand. Other types of models require more effort to make sense of their performance on specific inputs. Some non-inherently interpretable models can be made interpretable in a post-hoc fashion through explanation or visualization. Model-agnostic post-hoc methods can help to make models intelligible without custom explanation or visualization technologies and without changing the underlying model to make them more interpretable [37, 43].

Explanation generation can be described as a form of *post-hoc interpretability* [27, 31]; explanations are generated on-demand based on the current state of a model and—potentially—meta-knowledge about how the algorithm works. An important distinction between interpretability and explanation is that explanation does not elucidate precisely how a model works but aims to give useful information for practitioners and end users. Abdul et al. [2] conduct a comprehensive survey on trends in explainable and intelligible systems research.

Our work on rationale generation is a model-agnostic explanation system that works by translating the internal state and action representations of an arbitrary reinforcement learning system into natural language. Andreas, Dragan, and Klein [3] describe a technique that translates message-passing policies between two agents into natural language. An alternative approach to translating internal system representations into natural language is to add explanations to a supervised training set such that a model learns to output a classification as well as an explanation [13]. This technique has been applied to generating explanations about procedurally generated game level designs [21].

Beyond the technology, user perception and acceptance matter because they influence trust in the system, which is crucial to adoption of the technology. Established fields such as information systems enjoy a robust array of technology acceptance models such as the Technology Acceptance Model (TAM) [14] and Unified Theory of Acceptance and Use of Technology Model (UTAUT) [39]

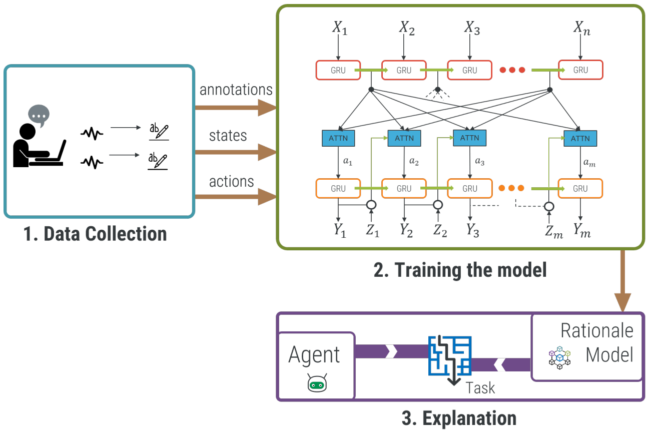
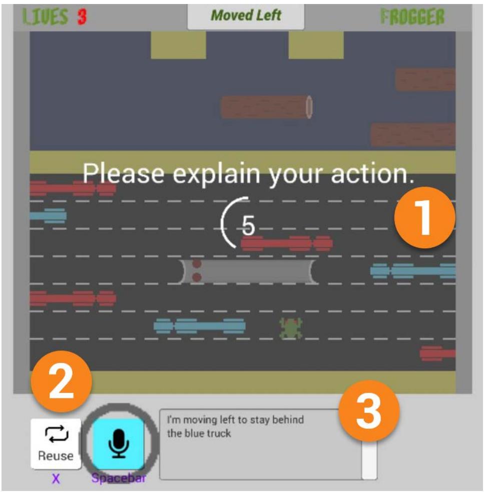


Figure 1: End-to-end pipeline for training a system that can generate explanations.



I'm moving left to stay behind the blue truck

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 if needed.

whose main goal is to explain variables that influence user perceptions. Utilizing dimensions such as perceived usefulness and perceived ease of use, the TAM model aimed to explain prospective expectations about the technological artifacts. UTAUT uses constructs like performance expectancy, effort expectancy, etc. to understand technology acceptance. The constructs and measures in these models build on each other.

In contrast, due to a rapidly evolving domain, a robust and well-accepted user perception model of XAI agents is yet to be developed. Until then, we can take inspiration from general acceptance models (such as TAM and UTAUT) and adapt their constructs to understand the perceptions of XAI agents. For instance, the human-robot interaction community has used them as basis to understand users' perceptions towards robots [6, 18]. While these acceptance models are informative, they often lack sociability factors such as "humanlike-ness". Moreover, TAM-like models does not account for autonomy in systems, let alone autonomous XAI systems. Building on some constructs from TAM-like models and original formative work, we attempt to address the gaps in understanding user perceptions of rationale-generating XAI agents.

The dearth of established methods combined with the variable conceptions of explanations make evaluation of XAI systems challenging. Binns et al. [7] use scenario-based survey design [11] and presented different types of hypothetical explanations for the same decision to measure perceived levels of justice. One non-neural based network evaluates the usefulness and naturalness of generated explanations [10]. Rader et al. [36] use explanations manually generated from content analysis of Facebook's News Feed to study perceptions of algorithmic transparency. One key differentiating factor of our approach is that our evaluation is based rationales that are actual system outputs (compared to hypothetical ones). Moreover, user perceptions of our system's rationales directly influence the design of our rationale generation technique.

3 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 necessarily 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).

We experiment with rationale generation using autonomous agents that play the arcade game, *Frogger*. 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, making it an ideal stepping stone towards a real world environment. Our rationale generation technique is agnostic to the type of agent or how it is trained, as long as the representations of states and actions used by the agent can be exposed to the rationale generator and serialized.

3.1 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 a rationale-generating explainable agent, 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 (1), 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.

For rationale collection (2), participants play through the game while explaining their actions out loud in a turn-taking mechanism. Figure 2 shows the game embedded into the explanation collection interface. To help couple explanations with actions (attach annotations to concrete game states), the game pauses for 10 seconds after each 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 [1] automatically transcribes the explanation real-time. The automatic transcription substantially reduces participant burden as it is more efficient than typing an explanation. Player can use more or less than the default 10-second pause to collect the explanation. Once done explaining, they can view their transcribed text and edit it if necessary. During pretesting with 14 players, we observed that players often repeat a move for which the explanation is the same as before. To reduce burden of repetition, we added a "redo" button that can be used to recycle rationales for consecutive repeated actions.

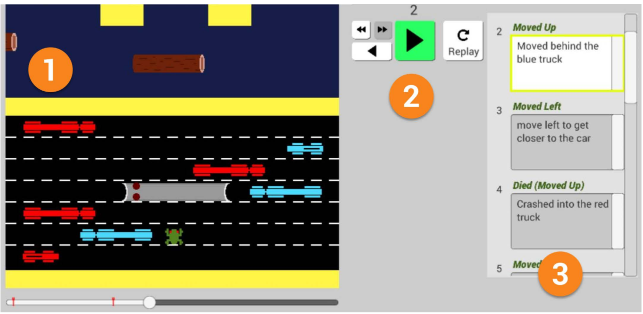
When the game play is over, players move to transcribed explanation review portion (3). Here, they can can step through all the actions-explanation pairs. This stage allows reviewing in both a situated and global context.

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

3.2 Neural Translation Model

We use an encoder-decoder network [5] 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 our task. Our encoder decoder architecture is similar to that used in [17]. The network learns how to translate the input game state representation $X = x_1, x_2, ..., x_n$, comprised of the representation of the game combined with other influencing factors, into an output rationale as a sequence of words $Y = y_1, y_2, ..., y_m$ where y_i is a word. 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 Gated Recurrent Unit (GRU) cells since our training process involved a small amount of data. The decoder network uses an additional attention mechanism [29] to learn to weight the importance of different components of the input with regard to their effect on the output.



2 Moved Up
Moved behind the blue truck

3 Moved Left
move left to get closer to the car

4 Died (Moved Up)
Crashed into the red truck

5 Moved

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

To simplify the learning process, the state of the game environment is serialized into a sequence of symbols where each symbol characterizes a sprite in the grid-based represntation of the world. 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. Empirically, we found that a reinforcement learning agent using tabular Q-learning [41] learns to play the game effectively when given a limited window for observation. Thus a natural configuration for the rationale generator is to give it the same observation window that the agent needs to learn to play. We refer to this configuration of the rationale generator as *focused-view* generator. This view, however, potentially limits the types of rationales that can be learned since the agent will only be able to see a subset of the full state. Thus we formulated a second configuration that gives the rationale generator the ability to use all information on the board to produce rationales. We refer to this as *complete-view* generator. An underlying question is thus whether rationale generation should use the same information that the underlying black box reasoner needs to solve a problem or if more information is advantageous at the expense of making rationale generation a harder problem. In the studies described below, we seek to understand how these configurations affect human perceptions of the agent when presented with generated rationales.

3.2.1 Focused-view Configuration. In the *focused-view* 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 designed the inputs such that the network is prone to prioritize short-term planning producing localized rationales instead of long-term planning.

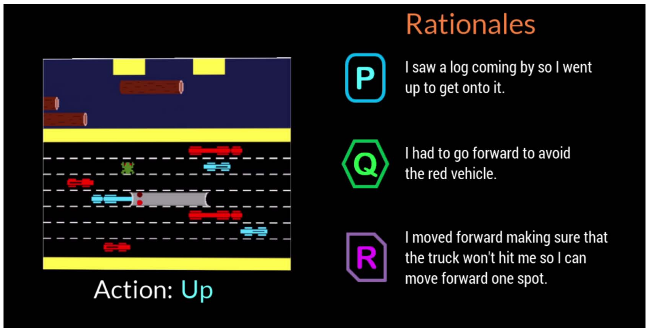
3.2.2 Complete-view Configuration. The *complete-view* configuration is an alternate setup that provides the entire game board as context for the rationale generation. There are two differences between this configuration and the focused-view configuration. First, we use the entire game screen as a part 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, reduce the likelihood that spurious correlations are identified, and to 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. Given the input structure and scope, this configuration should prioritize rationales that exhibit long-term planning and consider the broader context.

Table 1: Examples of *focused-view* vs *complete-view* rationales generated by our system for the same set of actions.

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.

4 PERCEPTION STUDY: CANDIDATE VS. BASELINE RATIONALES

In this section, we assess whether the rationales generated using our technique are plausible and explore how humans perceive them along various dimensions of human factors. For our rationales to be plausible we would expect that human users indicate a strong preference for rationales generated by our system (either configuration) over those generated by a baseline rationale generator. We also compare them to exemplary human-produced explanations to get a sense for how far from the upper bound we are.



Rationales

P I saw a log coming by so I went up to get onto it.

Q I had to go forward to avoid the red vehicle.

R I moved forward making sure that the truck won't hit me so I can move forward one spot.

Action: Up

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

This study aims to achieve two main objectives. First, it seeks to confirm the hypothesis that humans prefer rationales generated by each of the configurations over randomly selected rationales across all dimensions. While this baseline is low, it establishes that rationales generated by our technique are not nonsensical. We can also measure the distance from the upper-bound (exemplary human rationales) for each rationale type. Second, we attempt to understand the underlying components that influence the perceptions of the generated rationales along four dimensions of human factors: *confidence*, *human-likeness*, *adequate justification*, and *understandability*.

4.1 Method

To gather the training set of game state annotations, we deployed our data collection pipeline on *Turk Prime* [28]. From 60 participants we collected over 2000 samples of human actions in Frogger coupled with natural language explanations. The average duration of this task was around 36 minutes. The parallel corpus of the collected game state images and natural language explanations 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.

For the perception user study, we collected both within-subject and between-subject data. We recruited 128 participants, split into two equal experimental groups through *TurkPrime*: Group 1 (age range = 23 - 68 years, M = 37.4, SD = 9.92) and Group 2 (age range = 24 - 59 years, M = 35.8, SD= 7.67). On average, the task duration was approximately 49 minutes. 46% of our participants were women, and the 93% of participants were self-reported as from the United States while the remaining 7% of participants were self-reported as from India.

All participants watched a counterbalanced series of five videos. Each video depicted an action taken by Frogger accompanied by three different types of rationales that justified the action (see Figure 4). Participants rated each rationale on a labeled, 5-point, bipolar Likert-scale along 4 perception dimensions (described below). Thus, each participant provided 12 ratings per action, leading to 60 perception ratings for five actions. Actions collected from human players comprised the set of Frogger's actions. These actions were then fed into the system to generate rationales to be evaluated in the the user studies. In order to get a balance between participant burden, fatigue, the number of actions, and regions of the game, we pretested with 12 participants. Five actions was the limit beyond which participants' fatigue and burden substantially increased. Therefore, we settled on five actions (up (twice), down, left, and right) in the major regions of the game– amongst the cars, at a transition point, and amongst the logs. This allowed us to test our rationale generation configurations in all possible action-directions in all the major sections of the game.

The study had two identical experimental conditions, differing only by type of *candidate rationale*. Group 1 evaluated the *focused-view* rationale while Group 2 evaluated the *complete-view* rationales. 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 until consensus was reached. This is provided as an upper-bound for contrast with the next two techniques.
- The *candidate rationale* is the rationale produced by our network, either the focused-view or complete-view configuration.
- 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.

1. *Confidence:* This rationale makes me confident in the character's ability to perform it's task.
2. *Human-likeness:* This rationale looks like it was made by a human.
3. *Adequate justification:* This rationale adequately justifies the action taken.
4. *Understandability:* This rationale helped me understand why the agent behaved as it did.

Response options on a clearly labeled bipolar Likert scale ranged from "strongly disagree" to "strongly agree". In a mandatory free-text field, they explained their reasoning behind the ratings for a particular set of three rationales. After answering these questions, they provided demographic information.

These four dimensions emerged from an iterative filtering process that included preliminary testing of the study, informal interviews with experts and participants, and a literature review on robot and technology acceptance models. Inspired by the acceptance models, we created a set of dimensions that were contextually appropriate for our purposes.

Direct one-to-one mapping from existing models was not feasible, given the novelty and context of the Explainable AI technology. We adapted *confidence*, a dimension that impacts trust in the system [23], from constructs like performance expectancy [39] (from UTAUT) and robot performance [6, 12]. *Human-likeness*, central to generating human-centered rationales, was inspired from sociability and anthropomorphization factors from HRI work on robot acceptance [[33–35]. Since our rationales are justificatory in nature, *adequate justification* is a reasonable measure of output quality (transformed from TAM). Our rationales also need to be *understandable*, which can signal perceived ease of use (from TAM).

4.2 Quantitative Analysis

We used a multi-level model to analyze our data. All variables were within-subjects except for one: whether the candidate style was focused-view (Group 1) or complete-view (Group 2). This was a between-subject variable.

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 ($b = 1.90, t(252) = 8.09, p < .001$). In addition, exemplary rationales were judged significantly higher than candidate rationales.

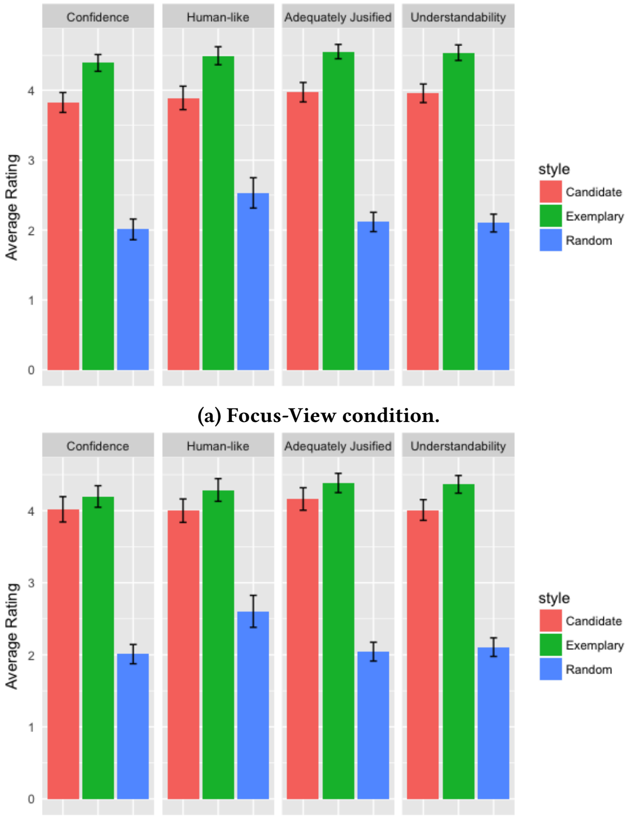
Though there were significant differences between each kind of candidate rationale and the exemplary rationales, those differences were not the same. The difference between the *focused-view* candidate rationales and exemplary rationales were significantly *greater* than the difference between *complete-view* 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 higher ratings in general than did the *focused-view* candidate rationales ($t(1530) = 8.33, p < .001$).

In summary, we have confirmed our hypothesis that both configurations produce rationales that perform significantly better than the *random* baseline across all dimensions.

4.3 Qualitative Findings and Discussion

In this section, we look at the open-ended responses provided by our participants to better understand the 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 [38]. 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 five most relevant components: (1) *Contextual Accuracy*, (2) *Intelligibility*, (3) *Awareness*, (4) *Relatability*, and (5) *Strategic Detail* (see Table 3). 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.



(a) Focus-View condition.

(b) Complete-View condition.

Figure 5: Human judgment results.

We will now share how these components influence the dimensions of the human factors under investigation. When providing examples of our participants’ responses, we will use P1 to refer to participant 1, P2 for participant 2, etc. To avoid priming during evaluation, we used letters (e.g., A, B, C, etc.) to refer to the different types of rationales. For better comprehension, we have substituted the letters with appropriate rationale–focused-view, complete-view, or random– while presenting quotes from participants below.

4.3.1 Confidence (1). 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, 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.

Table 2: 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.

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 because they refer to an object not present or wrong time sequence, so I had very low confidence. (P66)

4.3.2 Human-likeness (2). 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 as it depended on what participants thought “being human” entailed. Some conceptualized humans as fallible beings 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 “. . . [laid] 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 detail 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.

4.3.3 Adequate Justification (3). 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* in the form of irrelevant details. As P11 puts it:

> The [random rationale] doesn'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 score 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 rated *exemplary* and *focused-view* rationales high because each made the participant "believe in the character's ability to judge their surroundings."

4.3.4 Understandability (4). 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. In terms of *contextual accuracy*, many expressed how the contextual accuracy, not the length of the rationale, mattered when it came to understandability. While comparing understandability of the *exemplary* and *focused-view* rationales, 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 rationale (typically lengthier than focused)] was not as applicable because the [focused-view] rationale does a better job of providing contextual 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.

In summary, the first study establishes the plausibility of the generated rationales (compared to baselines) and their user perceptions. However, this study does not provide direct comparison between the two configurations.

5 PREFERENCE STUDY: FOCUSED– VS. COMPLETE–VIEW RATIONALES

The preference study puts the rationales in direct comparison with each other. This study It achieves two main purposes. First, it aims to validate the alignment between the intended design of rationale types and the actual perceived differences between them. We collect qualitative data on how participants perceived rationales produced by our *focused-view* and *complete-view* rationale generator. Our expert observation is that the *focused-view* configuration results in concise and localized rationales whereas the *complete-view* configuration results in detailed, holistic rationales. We seek to determine whether naïve users who are unaware of which configuration produced a rationale also describe the rationales in this way. Second, we seek to understand how and why the preferences between the two styles differed along three dimensions: *confidence*, *failure*, and *unexpected behavior*.

5.1 Method

Using similar methods to the first study, we recruited and analyzed the data from 65 people (age range = 23 - 59 years, M = 38.48, SD = 10.16). 57% percent of the participants were women with 96% of the participants self-reporting the United States and 4% self-reporting India as countries they were from. Participants from our first study could not partake in the second one. The average task duration was approximately 46 minutes.

The only difference in the experimental setup between perception and the preference study is the comparison groups of the rationales. In this study, participants judged the same set of *focused*- and *complete-view* rationales, however instead of judging each style against two baselines, participants evaluate the *focused*- and *complete-view* rationales in direction comparison with each other.

Having watched the videos and accompanying rationales, participants responded to the following questions comparing both configurations:

1. **Most important difference**: What do you see as the most important difference? Why is this difference important to you?
2. **Confidence**: Which style of rationale makes you more confident in the agent's ability to do its task? Was it system A or system B? Why?
3. **Failure**: If you had a companion robot that had just made a mistake, would you prefer that it provide rationales like System A or System B? Why?
4. **Unexpected Behaviour**: If you had a companion robot that took an action that was not wrong, but unexpected from your perspective, would you prefer that it provides rationales like System A or System B? Why?

We used a similar to the process of selecting dimensions in this study as we did in the first one. *Confidence* is crucial to trust especially when failure and unexpected behavior happens [12, 23]. Collaboration, tolerance, and perceived intelligence are affected by the way autonomous agents and robots communicate *failure* and *unexpected behavior* [15, 24, 25, 32].

Table 3: Tally of how many preferred the *focused-view* vs. the *complete-view* for the three dimensions.

Question	Focused-view	Complete-view
Confidence	15	48
Failure	17	46
Unexpected Behaviour	18	45

5.2 Quantitative Analysis

In order to determine whether the preferences significantly favored one style or the other, we conducted the Wilcoxon signed-rank test. It showed that preference for the *complete-view* rationale was significant in all three dimensions. Confidence in the *complete-view* rationale was significantly greater than in the *focused-view* ($p < .001$). Similarly, preference for a *complete-view* rationales from an agent that made a mistake was significantly greater than for *focused-view* rationales ($p < .001$). Preference for *complete-view* rationales from an agent that made a mistake was also significantly greater than for *focused-view* rationales ($p < .001$).

5.3 Qualitative Findings and Discussion

In this section, similar to the first study, we share insights gained from the open-ended responses to reveal the underlying reasons behind perceptions of the *most important difference* between the two styles. We also unpack the reasoning behind the quantitative ranking preferences for *confidence* in the agent's ability to do its task and communication preferences for *failure* and *unexpected behavior*. In this analysis, the interacting *components* that influenced the dimensions of human factors in the first study return (see Table 3). In particular, we use them as analytic lenses to highlight the trade-offs people make when expressing their preferences and the reasons for the perceived differences between the styles.

These insights bolster our situated understanding of the differences between the two rationale generation techniques and assist to verify if the intended design of the two configurations aligns with the perceptions of them. In essence, did the design succeed in doing what we set out to do? We analyzed the open-ended responses in the same manner as the first study. We use the same nomenclature to refer to participants.

5.3.1 Most Important Difference (1). Every participant indicted that the *level of detail and clarity* (P55) differentiated the rationales. Connected to the level of detail and clarity is the perceived *long*- vs. *short-term* planning exhibited by each rationale. Overall, participants felt that the *complete-view* rationale showed better levels of *strategic detail*, *awareness*, and *relatability* with human-like justifications, whereas the *focused-view* exhibited better *intelligibility* with easy-to-understand rationales. The following quote illustrates the trade-off between succinctness, which hampers comprehension of higher-order goals, and broadness, which can be perceived as less focused:

> The [focused-view rationale] is extraordinarily vague and focused on the raw mechanics of the very next move ... [The complete-view] is more broad and less focused, but takes into account *the entire picture.* So I would say the most important difference is the *scope of events* that they take into account while making justifications [emphasis added] (P24)

Beyond trade-offs, this quote highlights a powerful validating point: without any knowledge beyond what is shown on the video, the participant pointed out how the *complete-view* rationale appeared to consider the "entire picture" and how the "scope of events" taken into account was the main difference. The participant's intuition precisely aligns with the underlying network configuration design and our research intuitions. Recall that the *complete-view* rationale was generated using the entire environment or "picture" whereas the *focused-view* was generated using a windowed input.

In prior sections, we speculated on the effects of the network configurations. We expected the *focused-view* version to produce succinct, localized rationales that concentrated on the short-term. We expected the *complete-view* version to produce detailed, broader rationales that focused on the larger picture and long-term planning. The findings of this experiment are the first validation that the outputs reflect the intended designs. The strength of this validation was enhanced by the many descriptions of our intended attributes, given in free-form by participants who were naive to our network designs.

Connected to the level of detail and clarity is the perception of *short*- vs *long-term* thinking from the respective rationales. In general, participants regarded the *focused-view* rationale having low levels of *awareness* and *strategic detail*. They felt that this agent "...focus[ed] only on the current step" (P44), which was perceived depicting as thinking "...in the spur of the moment" (P27), giving the perception of short-term and simplistic thinking. On the other hand, the *complete-view* rationale appeared to "...try to think it through" (P27), exhibiting long-term thinking as it appears to "...think forward to broader strategic concerns."(P65) One participant sums it up nicely:

> The [focused-view rationale] focused on the immediate action required. [The complete-view rationale] took into account the current situation, [but] also factored in what the next move will be and what dangers that move poses. The [focused-view] was more of a short term decision and [complete-view] focused on both short term and long term goals and objectives. (P47)

We will notice how these differences in perception impact other dimensions such as confidence and communication preferences for failure and unexpected behavior.

5.3.2 Confidence (2). Participants had more confidence in the agent's ability to do its task if the rationales exhibited high levels of *strategic detail* in the form of long-term planning, *awareness* via expressing knowledge of the environment, and *relatability* through humanlike expressions. They associated *conciseness* with confidence when the rationales did not need to be detailed given the context of the (trivial) action.

The *complete-view* rationale inspired more confidence because participants perceived agents with long-term planning and high *strategic detail* as being "more predictive" and intelligent than their counterparts. Participants felt more at ease because "...knowing what [the agent] was planning to do ahead of time would allow me to catch mistakes earlier before it makes them." (P31) As one participant put it:

> The [complete-view rationale] gives me more confidence ... because it thinks about future steps and not just the steps you need to take in the moment. [The agent with focused-view] thinks more simply and is prone to mistakes. (P13)

Participants felt that rationales that exhibited a better understanding of the environment, and thereby better *awareness*, resulted in higher confidence scores. Unlike the *focused-view* rationale that came across as "a simple reactionary move ... [the *complete-view*] version demonstrated a more thorough understanding of the entire field of play." (P51) In addition, the *complete-view* was more *relatable* and confidence-inspiring "because it more closely resemble[d] human judgment" (P29).

5.3.3 Failure (3). When an agent or a robot fails, the information from the failure report is mainly used to fix the issue. To build a mental model of the agent, participants preferred *detailed* rationales with solid *explanatory power* stemming from *awareness* and *relatability*. The mental model could facilitate proactive and preventative care.

The *complete-view* rationale, due to relatively high *strategic detail*, was preferable in communicating failure because participants could "...understand the full reasoning behind the movements."(P16) Interestingly, *detail* trumped *intelligibility* in most circumstances. Even if the rationales had some grammatical errors or were a "...little less easy to read, the details made up for it." (P62)

However, detailed rationales are not always a virtue. Simple rationales have the benefit of being easily understandable to humans, even if they cause humans to view the agent as having limited understanding capabilities. Some participants appreciated *focused-view* rationales because they felt "it would be easier to figure out what went wrong by focusing on one step at a time."

Explanatory power, specifically how events are communicated, is related to *awareness* and *relatability*. Participants preferred relatable agents that "...would talk to [them] like a person would."(P11) They expressed the need to develop a mental model, especially to "...see how [a robot's] mind might be working"(P1), to effectively fix the issue. The following participant neatly summarizes the dynamics:

> I'd want [the robot with complete-view] because I'd have a better sense of the steps taken that lead to the mistake. I could then fix a problem within that reasoning to hopefully avoid future mistakes. The [focused-view rationale] was just too basic and didn't give enough detail. (P8)

5.3.4 Unexpected Behavior (4). Unexpected behavior that is not failure makes people want to know the "why?" behind the action, especially to understand the expectancy violation. As a result, participants preferred rationales with transparency so that they can understand and trust the robot in a situation where expectations are violated. In general, preference was for adequate levels of *detail* and *explanatory power* that could provide "...more diagnostic information and insight into the robot's thinking processes."(P19) Participants wanted to develop mental models of the robots so they could understand the world from the robot's perspective. This diagnostic motivation for a mental model is different from the reprogramming or fixing needs in cases of failure.

The *complete-view* rationale, due to adequate levels of *strategic detail*, made participants more confident in their ability to follow the thought process and get a better understanding of the expectancy violation. One participant shared:

> The greater clarity of thought in the [complete-view] rationale provides a more thorough picture ..., so that the cause of the unexpected action could be identified and explained more easily. (P51)

With this said, where possible without sacrificing transparency, participants welcomed simple rationales that "anyone could understand, no matter what their level of education was."(P2) This is noteworthy because the expertness level of the audience is a key concern when making accessible AI-powered technology where designers need to strike a balance between detail and succinctness.

Rationales exhibiting strong explanatory power, through *awareness* and *relatability*, helps to situate the unexpected behavior in an understandable manner. Participants preferred the *complete-view* rationale's style of communication because of increased transparency:

> I prefer [the complete-view rationale style] because ...I am able to get a much better picture of why it is making those decisions. (P24)

Despite similarities in the communication preferences for failure and unexpected behavior, there are differences in underlying reasons. As our analysis suggests, the mental models are desired in both cases, but for different reasons.

6 DESIGN LESSONS AND IMPLICATIONS

The situated understanding of the *components* and *dimensions* give us a powerful set of actionable insights that can help us design better human-centered, rationale-generating, autonomous agents. As our analysis reveals, context is king. Depending on the context, we can tweak the input type to generate *rationale sytles* that meet the needs of the task or agent persona; for instance, a companion agent that requires high *relatability* for user engagement. We should be mindful when optimizing for a certain dimension as each component comes with costs. For instance, conciseness can improve *intelligibility* and overall *understandability* but comes at the cost of *strategic detail*, which can hurt *confidence* in the agent. We can also engineer systems such that multiple network configurations act as modules. For instance, if we design a companion agent or robot that interacts with a person longitudinally, the *focused-view* configuration can take over when short and simple rationales are required. The *complete-view* configuration or a hybrid one can be activated when communicating failure or unexpected behavior.

As our preference study shows, we should not only be cognizant about the level of detail, but also why the detail is necessary, especially while communicating failure and unexpected behavior. For instance, failure-reporting, in a mission critical task (such as search and rescue), would have different requirements for *strategic detail* and *awareness*, compared to "failure" reporting in a less-defined, more creative task like making music. While the focus of this paper is on textual rationale generation, rationales can be complementary to other types of explanations; for instance, a multi-modal system can combine visual cues with textual rationales to provide better contextual explanations for an agent's actions.

7 LIMITATIONS AND FUTURE WORK

While these results are promising, there are several limitations in our approach that need to be addressed in future work. First, our current system, by intention and design, lacks interactivity; users cannot contest a rationale or ask the agent to explain in a different way. To a get a formative understanding, we kept the design as straight-forward as possible. Now that we have a baseline understanding, we can vary along the dimension of interactivity for the next iteration. For instance, contestability, the ability to either reject a reasoning or ask for another one, which has shown to improve user satisfactions [16, 22] can be incorporated in the future. Second, 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. In continuous-time and -action environments, we must determine how to collect the necessary data without being too intrusive to participants. Third, all conclusions about our approach were formed based on one-time interactions with the system. To better control for potential novelty effects that rationales could have, we need to deploy our system in a longitudinal task setting. Fourth, to understand the feasibility of our system in larger state-action spaces, we would need to study the scalability by addressing the question of how much data is needed based on the size of environment. Fifth, not all mistakes are created equal. Currently, the perception ratings are averaged where everything is equally weighted. For instance, a mistake during a mission critical step can lead to higher fall in confidence than the same mistake during a non-critical step. To understand the relative costs of mistakes, we need to further investigate the relationship between context of the task and the cost of the mistake.

8 CONCLUSIONS

While explainability has been successfully introduced for classification and captioning tasks, sequential environments offer a unique challenge for generating human understandable explanations. The challenge stems from multiple complex factors, such as temporally connected decision-making, that contribute to making decisions in these environments. In this paper, we introduce *automated rationale generation* as a concept and explore how justificatory explanations from humans can be used to train systems to produce human-like explanations in sequential environments. To facilitate this work, we also introduce a pipeline for automatically gathering a parallel corpus of states annotated with human explanations. This tool enables us to systematically gather high quality data for training purposes. We then use this data to train a model that uses machine translation technology to generate human-like rationales in the arcade game, *Frogger*.

Through a mixed-methods approach in evaluation, we establish the plausibility of the generated rationales and describe how intended design of rationale types lines up with the actual user perceptions of them. We also get contextual understanding of the underlying dimensions and components that influence human perception and preferences of the generated rationales. By enabling autonomous agents to communicate about the motivations for their actions, we envision a future where explainability not only improves human-AI collaboration, but does so in a human–centered and understandable manner.

9 ACKNOWLEDGEMENTS

This work was partially funded under ONR grant number N00014141000. We would like to thank Chenghann Gan and Jiahong Sun for their valuable contributions to the development of the data collection pipeline. We are also grateful to the feedback from anonymous reviewers that helped us improve the quality of the work.

REFERENCES

[1] 2017. streamproc/MediaStreamRecorder. (Aug 2017). https://github.com/streamproc/MediaStreamRecorder

[2] 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*. ACM, 582.

[3] Jacob Andreas, Anca Dragan, and Dan Klein. 2017. Translating neuralese. *arXiv preprint arXiv:1704.06960* (2017).

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

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[6] Jenay M Beer, Akanksha Prakash, Tracy L Mitzner, and Wendy A Rogers. 2011. *Understanding robot acceptance.* Technical Report. Georgia Institute of Technology.

[7] 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.

[8] Ned Block. 2005. Two neural correlates of consciousness. *Trends in cognitive sciences* 9, 2 (2005), 46–52.

[9] Ned Block. 2007. Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and brain sciences* 30, 5-6 (2007), 481–499.

[10] Joost Broekens, Maaike Harbers, Koen Hindriks, Karel Van Den Bosch, Catholijn Jonker, and John-Jules Meyer. 2010. Do you get it? User-evaluated explainable BDI agents. In *German Conference on Multiagent System Technologies*. Springer, 28–39.

[11] John M Carroll. 2000. *Making use: scenario-based design of human-computer interactions.* MIT press.

[12] Sonia Chernova and Manuela M Veloso. 2009. A Confidence-Based Approach to Multi-Robot Learning from Demonstration.. In *AAAI Spring Symposium: Agents that Learn from Human Teachers*. 20–27.

[13] Noel CF Codella, Michael Hind, Karthikeyan Natesan Ramamurthy, Murray Campbell, Amit Dhurandhar, Kush R Varshney, Dennis Wei, and Aleksandra Mojsilovic. 2018. Teaching Meaningful Explanations. *arXiv preprint arXiv:1805.11648* (2018).

[14] Fred D Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly* (1989), 319–340.

[15] Munjal Desai, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. 2013. Impact of robot failures and feedback on real-time trust. In *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 251–258.

[16] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2016. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64, 3 (2016), 1155–1170.

[17] Upol Ehsan, Brent Harrison, Larry Chan, and Mark O. Riedl. 2018. Rationalization: A Neural Machine Translation Approach to Generating Natural Language Explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence, Ethics, and Society*.

[18] Neta Ezer, Arthur D Fisk, and Wendy A Rogers. 2009. Attitudinal and intentional acceptance of domestic robots by younger and older adults. In *International Conference on Universal Access in Human-Computer Interaction*. Springer, 39–48.

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

[20] Marsha E Fonteyn, Benjamin Kuipers, and Susan J Grobe. 1993. A description of think aloud method and protocol analysis. *Qualitative health research* 3, 4 (1993), 430–441.

[21] Matthew Guzdial, Joshua Reno, Jonathan Chen, Gillian Smith, and Mark Riedl. 2018. Explainable PCGML via Game Design Patterns. *arXiv preprint arXiv:1809.09419* (2018).

[22] Tad Hirsch, Kritzia Merced, Shrikanth Narayanan, Zac E Imel, and David C Atkins. 2017. Designing contestability: Interaction design, machine learning, and mental health. In *Proceedings of the 2017 Conference on Designing Interactive Systems*. ACM, 95–99.

[23] Poornima Kaniarasu, Aaron Steinfeld, Munjal Desai, and Holly Yanco. 2013. Robot confidence and trust alignment. In *Human-Robot Interaction (HRI), 2013 8th ACM/IEEE International Conference on*. IEEE, 155–156.

[24] Minae Kwon, Sandy H Huang, and Anca D Dragan. 2018. Expressing Robot Incapability. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 87–95.

[25] Min Kyung Lee, Sara Kiesler, Jodi Forlizzi, Siddhartha Srinivasa, and Paul Rybski. 2010. Gracefully mitigating breakdowns in robotic services. In *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*. IEEE, 203–210.

[26] Peter Lipton. 2001. What good is an explanation? In *Explanation*. Springer, 43–59.

[27] Z. C. Lipton. 2016. The Mythos of Model Interpretability. *ArXiv e-prints* (June 2016).

[28] Leib Litman, Jonathan Robinson, and Tzvi Abberbock. 2017. TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior research methods* 49, 2 (2017), 433–442.

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

[30] Shadd Maruna and Ruth E Mann. 2006. A fundamental attribution error? Rethinking cognitive distortions. *Legal and Criminological Psychology* 11, 2 (2006), 155–177.

[31] Tim Miller. 2017. Explanation in artificial intelligence: insights from the social sciences. *arXiv preprint arXiv:1706.07269* (2017).

[32] Nicole Mirnig, Gerald Stollnberger, Markus Miksch, Susanne Stadler, Manuel Giuliani, and Manfred Tscheligi. 2017. To err is robot: How humans assess and act toward an erroneous social robot. *Frontiers in Robotics and AI* 4 (2017), 21.

[33] Clifford Nass, BJ Fogg, and Youngme Moon. 1996. Can computers be teammates? *International Journal of Human-Computer Studies* 45, 6 (1996), 669–678.

[34] Clifford Nass and Youngme Moon. 2000. Machines and mindlessness: Social responses to computers. *Journal of social issues* 56, 1 (2000), 81–103.

[35] Clifford Nass, Jonathan Steuer, Lisa Henriksen, and D Christopher Dryer. 1994. Machines, social attributions, and ethopoeia: Performance assessments of computers subsequent to" self-" or" other-" evaluations. *International Journal of Human-Computer Studies* 40, 3 (1994), 543–559.

[36] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 103.

[37] 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*. ACM, 1135–1144.

[38] Anselm Strauss and Juliet Corbin. 1994. Grounded theory methodology. *Handbook of qualitative research* 17 (1994), 273–85.

[39] 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.

[40] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. 2017. Residual Attention Network for Image Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.

[41] Christopher Watkins and Peter Dayan. 1992. Q-learning. *Machine learning* 8, 3-4 (1992), 279–292.

[42] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. 2048–2057.

[43] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579* (2015).

[44] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4651–4659.