

THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

DEPARTMENT OF MECHANICAL ENGINEERING

Determining Human Performance Given Robot-Provided Explanations Through Task-Based
Interaction

SYDNEY HANNAH
SPRING 2021

A thesis
submitted in partial fulfillment
of the requirements
for baccalaureate degrees
in Mechanical Engineering and International Politics
with honors in Mechanical Engineering

Reviewed and approved* by the following:

Alan Wagner
Assistant Professor of Aerospace Engineering
Thesis Supervisor

Bo Cheng
Associate Professor of Mechanical Engineering
Honors Adviser

Katie Fitzsimons
Assistant Professor of Mechanical Engineering
Faculty Reader

* Electronic approvals are on file.

ABSTRACT

As the applications and abilities of robots in society continue to progress, a thorough understanding of a robot's interaction with humans is critical. In practice, human-robot teams will need to communicate effectively in order to accomplish tasks and larger goals. Task-based goals that involve decision-making from one entity often require explanations to justify actions or ensure understanding among team members. This study investigates human performance based on the explanations provided by a robot that are intended to describe how the robot sorted a series of blocks and based on participant-created explanations. In this study, explanation is defined as communicating with the intention of describing information, which is in this case how a series of blocks was sorted into a pattern. Participants were guided by a robot in a simulation in which they were asked to identify the correct pattern based off the robot-provided explanation and rate quality of the robot's explanation in order to analyze participant performance. Participants were also asked to create their own pattern of blocks and explain it to the robot to measure matching of communication style.

Our findings indicate that participants were able to consistently distinguish between good, medium, and poor explanations, and rated the quality of explanations in the expected order. Surprisingly, we find that participants correctly identified patterns explained by the robot when the type of explanation provided was poor, and misidentify the pattern the most often when the type of explanation provided was of medium quality, which can be explained in part by the number of factors by which the patterns were sorted. Such findings cannot explicitly determine the relationship between human performance and quality of explanation, but still provide valuable insight into human performance in task-based interactions. Additionally, participant-

generated patterns and explanations offer insight on the convergence of linguistic and communication style and present ample opportunities for future work in human-robot communication on a psychological level. The findings from this experiment present important considerations for future research and utilization of robots in applied settings in which comprehensible explanation is pertinent.

TABLE OF CONTENTS

LIST OF FIGURES	iii
LIST OF TABLES	iv
ACKNOWLEDGEMENTS	v
Chapter 1 Introduction	1
Chapter 2 Literature Review	4
Movement and Understanding in Human-Robot Interaction	5
Robot-Provided Explanation in Human-Robot Interaction	6
Chapter 3 Methods	11
Simulation Setup	11
Chapter 4 Results	20
Ratings of Robot-Provided Explanation	20
Pattern Identification	21
Communication and Linguistic Style	23
Chapter 5 Discussion	25
Ratings of Robot-Provided Explanation	25
Pattern Identification	25
Communication and Linguistic Style	28
Chapter 6 Conclusion	31

LIST OF FIGURES

Figure 1. Kyle the Robot.....	12
Figure 2. Instruction Screen.....	13
Figure 3. Introduction and Practice Room.....	13
Figure 4. Survey Presented to Participants Following the Viewing of the Pattern and Robot's Explanation	15
Figure 5. Featured Pattern of Blocks Utilized in the Experiment, Sorted by Size.....	15
Figure 6. Participant View of Available Blocks for Pattern Creation.....	17
Figure 7. Example of a Drag-and-Drop Pattern.....	17
Figure 8. Chat Feature for Participants to Explain Pattern to the Robot.....	18
Figure 9. Participant Ratings of Explanations from 1 (very poor) to 5 (excellent).....	21
Figure 10. Correct Pattern Identification by Explanation Level	22
Figure 11. Participant Provided Explanations.....	24
Figure 12. Participant-Created Patterns	24

LIST OF TABLES

Table 1. Robot Behaviors	11
Table 2. Five Point Liker Scale Ratings of Explanations Presented by the Robot	20
Table 3. Pairwise Comparisons of Levels of Explanation	21
Table 4. Participant Identification of Patterns	22
Table 5. Robot-Provided Explanations	27

ACKNOWLEDGEMENTS

I would like to thank Dr. Alan Wagner for his guidance, support, and mentorship during this process. I would like to also thank Himavath Jois for being a constant sounding board and for working through an abundance of Unity glitches together. Thank you to all the members of the Robot Ethics and Aerial Vehicles Lab who helped me during this research, whether the assistance was with Unity, Amazon Mechanical Turk, or C#. Thank you to Dr. Carleen Maitland and Dr. Richard Canevez for showing me that research can and should be grounded in passion, and that research can truly make a positive impact on the world. Finally, thank you to my family for always being my strongest supporters and for encouraging me in all of my endeavors.

Chapter 1

Introduction

Explanation is a fundamental aspect of communication. Explanation allows for facilitated understanding of a topic, accomplishment of task-based goals, and deeper insight into a person.

As application and abilities of robots continue to progress, human-robot interaction and communication in everyday life will become more commonplace. Robot utilization will increasingly involve decision-making or explanation capabilities that will have a direct impact on a person. Even more important than a robot's ability to explain is a person's ability to apply the information contained within the robot's explanation such as through completion of a task.

Understanding how a person uses robot-provided explanations to perform tasks or answer related questions is valuable and provides guidance as to how to design robots in the future to maximize human performance of tasks. Such designs can vastly expand the applications of robots in various industries such as the medical or defense industries. Existent research has studied aspects of human-robot explanation including trust, adaptive behavior, and common ground [1]–[3]. This study aims to add to the literature by studying human performance given robot-provided explanations. This study also directs attention to the possibility of convergence or matching of communication styles between humans and robots, a phenomenon that is common in human-human communication.

This study first analyzes related works in the field to provide context for the experiment. The design and methods of the experiment are then expounded upon, explaining the specifics of

the interactive simulation created to test a series of hypotheses. These hypotheses were produced to address human performance given robot-provided explanations:

H1: Participants will rate the level of the robot's explanation higher for better patterns.

The notion behind this hypothesis is that explanations that are categorized by the researchers as better will also be categorized by participants as higher quality explanations. The accuracy of this hypothesis shapes the interpretation of the participants' responses to the rest of the experiment.

H2: Participants will be more likely to guess the correct pattern the better the explanation is. Quality of explanations were categorized with respect to the amount and quality of information contained within them, and it would track that a better explanation provided by the robot would result in higher performance in pattern identification among participants.

H3A: Participants will match their communication style with the robot when explaining their pattern to the robot. This hypothesis has the implication that participants will treat communication with and explaining something to a robot as similar or equivalent to communicating with or explaining something to another human.

H3B: Participants will prefer to explain their pattern in such a way that it matches the best explanation level the robot provided. Such a response would mean that participants are able to analyze the various levels of explanations provided by the robot and explain their own pattern matching the highest quality of explanation the robot used to explain to them.

In the simulation, a robot led the participant through a series of rooms as it presented pre-sorted patterns of blocks and explained the patterns to the participant. The participant was asked to rate the quality of the robot's explanations, to identify the correct pattern based off the robot's provided explanation, and to create their own pattern and explain it to the robot following a series of iterations of robot-provided explanations. The results from the experiment are then described and discussed in greater detail. Main findings from the experiment reveal that participants were able to consistently rate the quality of different levels of explanations, and that participants performance was not higher the better the robot-provided explanation was. An additional inconsistent component, the number of factors used to sort the pattern, which could include color, letter, size, or a combination of these factors, prevented determination of a connection between quality of explanation and participant performance. Participant generated explanations generally agreed with the robot-provided explanations, laying the groundwork for future experiments to study turn-taking and explanation-providing between a robot and participants over longer intervals in order to determine convergence over time. Applications of robots designed to maximize human performance are then discussed. Finally, future work and suggestions are outlined.

Chapter 2

Literature Review

Research by Wang et al. [4] focused on human to human interaction in a virtual setting and brought to light critical factors to consider when two entities, whether they are both human or a human and a robot, are conversing. In the study, participants either communicated just verbally, verbally and with haptics, or just with haptics. Haptics is a form of communication based on the sense of touch, just as speaking verbally is a form of communication based on the sense of sound. Results found that participants preferred to communicate verbally and that using just haptics to communicate will not be nearly as efficient between participants in a conversation where any information is being shared. In the scenario where both verbal communication and haptics were available to use during the interaction, verbal communication was heavily preferred over haptics. Such findings are imperative because they identify a more efficient and preferred mode of communication, task management, and explanation. Studies surrounding robots that have the ability to both speak and use haptics when communicating with a human can apply this knowledge to enhance interaction between a human and a robot to make it as fluid as possible.

Niederhoffer and Pennebaker [5] introduced a widely used method for analyzing communication known as linguistic style matching through a software package Pennebaker created known as Linguistic Inquiry and Word Count. This method analyzes similar utilization rates of certain words and parts of speech in conversation; it also produces scores that summarize the speakers' attitudes and self-assurance. They found that individuals in conversation converged in linguistic style as iterations or time engaged increased. An experiment that studied communication accommodation in instant messaging conversation between strangers and between friends found a convergence in message length and duration, but found less

convergence between strangers [6]. Studies such as these initially created for human to human interaction are critical to consider in human-robot interaction if the goal is to have a robot simulate the role of a human as closely as possible. Notably when it comes to verbal communication, convergence or divergence in speech acts as a way of interpreting what role an entity, either a robot or human, takes in conversation. Assessing utilization of certain types of words or attitudes while conversing also provides invaluable qualitative information to better understand how a human feels during the interaction. It is also imperative to keep in mind that while patterns can be determined from interactions, human explanation and understanding is subjective. As Keil [7] found, preferred types of explanations are individual and often based on preference. In essence, individuals may understand certain explanations from another better if they prefer that type of explanation; conversely, they may have a more difficult time understanding a conversation partner if the partner chooses to respond in a manner that is not the other individual's preferred method of explanation.

Movement and Understanding in Human-Robot Interaction

Following the study of verbal communication and adaptation in conversation between humans and artificial intelligence, studies surrounding movements and characteristics of robots during interaction became prominent. In a study conducted by Yamazaki *et al.* [8], a robot with a female voice guided the participant through a poster with information and shared facts about the subject being discussed. The robot's head moved during breaks in the conversation toward and away from the poster to which it referred. The study found that when the robot turned its head during a logical transition or break in the conversation, the participant was more likely to also

turn their head in coordination with the robot. In order to determine how the robot should move its head during conversation and while speaking, the researchers first studied human to human conversation. Head nods, gazes, and gestures helped determine the appropriate movements the robot should make during its interaction with a human. It is critical to study humans interacting with other humans because these movements can be applied to a robot, simulating similar body movement to humans in conversation. Replicating movements humans make during human to human interaction may help promote more fluid and effective conversations between robots and humans [9], [10].

Robot-Provided Explanation in Human-Robot Interaction

Recent work in the field of human-robot interaction reflects a shift in the role artificial intelligence plays as learning improves and robotics become more capable to perform a variety of tasks. An increase in robotic abilities expands the usage of robotics and industries in which robots will be implemented alongside humans to collaborate and perform tasks. Human performance based on robotic action and decision-making is a critical step in the field, the understanding of which will enhance robotic capabilities and applications in situations involving human contact. In a study conducted by Nikolaidis *et al.* [1], a robot and human attempted to move a table through a doorway in order to test human-robot teamwork and human response to robot instruction and suggestion. The findings proved that when it came to adapting behavior, individuals responded more positively when robots communicated with action statements and factual information rather than what is perceived by a human as reasoning or beliefs. Statements

uttered by the robot that humans recognized as self-doubt such as “I think” provided decreased trust levels of the robot regardless of its accuracy.

Studies have also concentrated on the human perception that robots have the capability to act with intention and explain that intention in ways that are comprehensible. Graaf and Malle [11] highlighted this need for robotic explanation capabilities to match those of humans in order to elevate human-robot interaction. Pulling from human psychology and human-robot interaction, the study dissected human tendencies to explain the conversation, their thought process, and their speech based on the context of the scenario. Robotic ability to follow this thought process and to apply it in conversation with a human enhances the interaction and allows the human to perceive the robot as a true participant. The ability to view a robot as an active participant with similar thought processes can have a positive impact on overall human perception of robots, but studies require a more sophisticated way of analyzing human perception of robots before and after the robot interaction in order to empirically understand these effects.

Nomura *et al.* [12] addressed the need for an applicable measurement scale with the creation of the Negative Attitudes Toward Robots Scale (NARS) and the Robot Anxiety Scale (RAS). Emotions play a role in human communication with others and will continue to play a role in communication with another entity, even if that entity is a robot in place of a human. The NARS and RAS tools represented the participant’s anxiety and overall uneasiness while interacting and communicating with a robot, choosing to focus on negative feelings toward robots rather than neutral or positive feelings. Applying a form of these scales to a study both before and after human-robot interaction is useful to assess prior participant assumptions and feelings about robots and the information gathered from these scales can have broad impacts on analysis of the study. Furthermore, these scales have aided the design of future iterations of

robots in experiments in order to decrease anxiety and willingness to communicate. A primary goal in human-robot interaction is to cultivate conversation equivalent to conversation held human to human. The work just discussed made strides in this area and allowed for focus on the more minute aspects of conversation that would enhance human-robot conversation, specifically creating a sense of common ground between the participants.

Kiesler [13] explicates how shared understanding and knowledge produce more efficient conversations. Humans tend to assign human attributes and characteristics to robots and will therefore respond similarly to a robot as they would to a human in conversation the more they understand the robot and its knowledge. Combining tools such as RAS with the theory posited by Kiesler will allow for the design of robots that have the ability to tailor how it presents itself to the human with which it interacts in order to create positive, efficient, and productive conversations.

Other studies in common ground discovered that humans use fewer words to explain topics of shared knowledge or background with the other conversation participant [2]. Powers *et al.* took this study a step farther into human-robot interaction by additionally assessing the types of responses of the participants prompted by the robot. To produce a longer and more thorough explanation from the participant, the robot appeared to have less understanding about the topic at hand, eliminating common ground from the conversation. An understanding of these human responses to robot queues provides a foundation to build upon in giving and receiving explanations between a robot and human in conversation.

Maaik *et al.* [14] expounded upon the preferences of explanation during conversation between a human and robot. The study, conducted specifically for the use of enhancing emergency preparedness training systems, asked trainers and instructors to measure the

usefulness of explanations provided by the AI system tasked with training them on a subject. By including instructors in this study, the effectiveness of the explanations provided by the system could be tailored to the specific task at hand, while also limiting the scope of the findings to situations that are similar to a teaching-based interaction. The instructors were first faced with a scenario and asked to give their own explanation that they would give to a trainee. After this, the instructors were asked to select the best explanation out of a bank of explanations provided by the AI system. The method applied here not only better the understanding of what types of explanations make the most sense for the emergency scenario, but also builds the bank of useful explanations that the AI can use, matching what actual instructors would use to explain to a trainee. The study found that being detailed and intentional in explanations is seen by instructors as favorable and indicates that the other participant in conversation will be more responsive to these types of explanations.

The quality of explanations often impacts trust. Wang *et al.* [3] analyzed the relationship between different types of explanation and trust. In the study, the participant and a robot were assigned as a team to check the safety of buildings in a city. Robots assigned to participants could assess the safety at either a high or low accuracy and not all the robots had the ability to explain the assessment. The robot could either explain using statistics, such as a confidence percentage, or observations. In this study, trust was measured by observing whether the participant listened to the robot. The study found that the participants trusted the robot more when the robot provided an explanation, even when it was less accurate. However, different types of explanations did not change the trust level of the participants. Participants reported that they understood the decision-making process of the robot even when only statistics were given and no observations were given. Another study developed a framework of trust based on

extensive studies into scenario-based trustworthiness and found that humans over-trusted robots even after mistakes were made [15]. This study directed attention to future research that explores more of the human dimension of human-robot trust and interaction, as this can provide critical insight into the field and advance future studies.

Other studies have explored methods for creating explanations in unfamiliar scenarios. Hanheide *et al.* [16] created and tested a method to explain failure of a task up to seven different ways. The robot explained task failure by combining its baseline knowledge and situation specific information. Their findings determined that the robot's explanation must include specific references to the individual scenario in order to properly explain the failure. This experiment did not evaluate the participants' understanding of the derived explanations. Determining how a robot should explain a task or failure to a human is critical, but just as critical is a human's ability to understand the explanation the robot is providing.

Chapter 3

Methods

Simulation Setup

The Unity simulation environment was used for these experiments. Unity allows developers to create simulations that can be experienced through web browsers. Unity was used to create a first-person simulation in which the participant could interact with a robot, follow a robot through rooms, answer questions, and create a unique pattern within the simulation. The necessary robot behaviors in the simulation were first defined, depicted in Table 1. The robot needed to be able to guide participants through the simulation and present patterns while providing explanations as to how the patterns were sorted. The final experimental design consisted of a series of rooms through which the robot guided each participant.

Table 1. Robot Behaviors

Required Behavior	Simulation Design
The robot must walk in fluid motion.	The robot was animated to walk as a human walks.
The robot must be aware of the objects in the simulation and not run into or through them.	The robot was programmed to have mass and navigate around objects in the simulation.
The robot must only walk to the next portion of the experiment when the participant is ready.	A series of interactive buttons and surveys appeared on the screen to guide the participant and indicated to the robot when it was appropriate to guide the participant into the next room.

Selection of the robot was influenced by the desire to make the simulation appear realistic and not overly similar to a video game. This led to the selection of the Unity Asset called Kyle the robot, pictured in Figure 2.

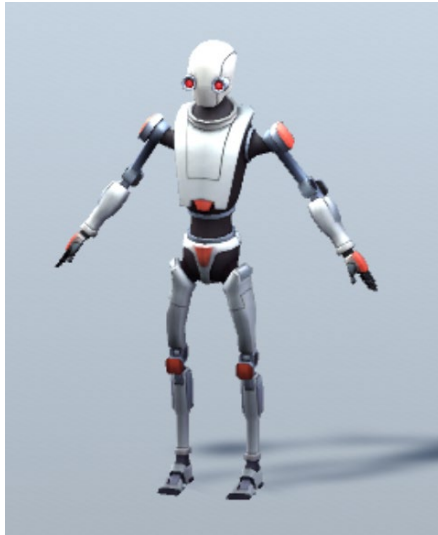


Figure 1. Kyle the Robot

The robot could be animated to walk, run, turn, or stand idle in organic, human-like motions, which encompassed the main requirements of the robot for this experiment. Available animations allowed the robot to guide participants to patterns, walk and stop when it arrived at its destinations, and interact with participants by facing them when it spoke via the chat function. Another important factor to consider during the design of the robot was the robot's linguistic style and disposition. Due to its functionality as a sorting and explanation-providing robot, the robot spoke in a formal tone. To maintain a formal syntax throughout the entirety of the experiment, the robot did not communicate using contractions. The robot's communication did not include extraneous information or conversational topics; its role was limited to guiding participants through the rooms and explaining its patterns.

Participants were recruited using Amazon Mechanical Turk and the entirety of the experiment was completed by participants through this platform. Amazon Mechanical Turk is a website in which researchers and businesses outsource tasks to be completed remotely by site-approved workers. A total of 27 participants completed this 10 minute experiment and were

compensated \$3.00. Participants for the experiment had an average age between 25-34 and were 85.2% male.

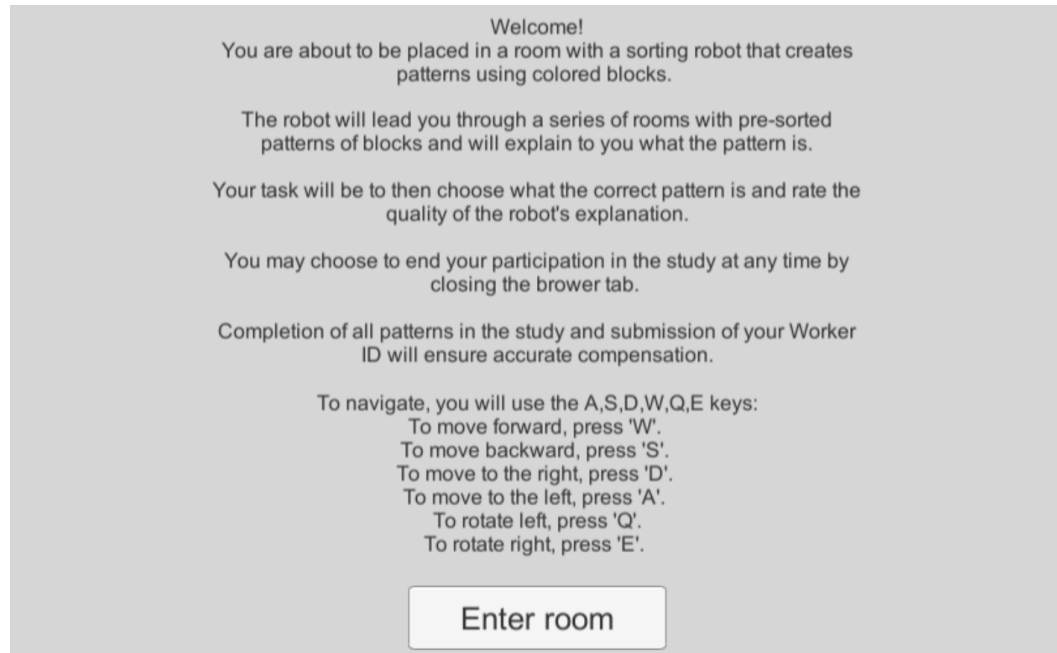


Figure 2. Instruction Screen

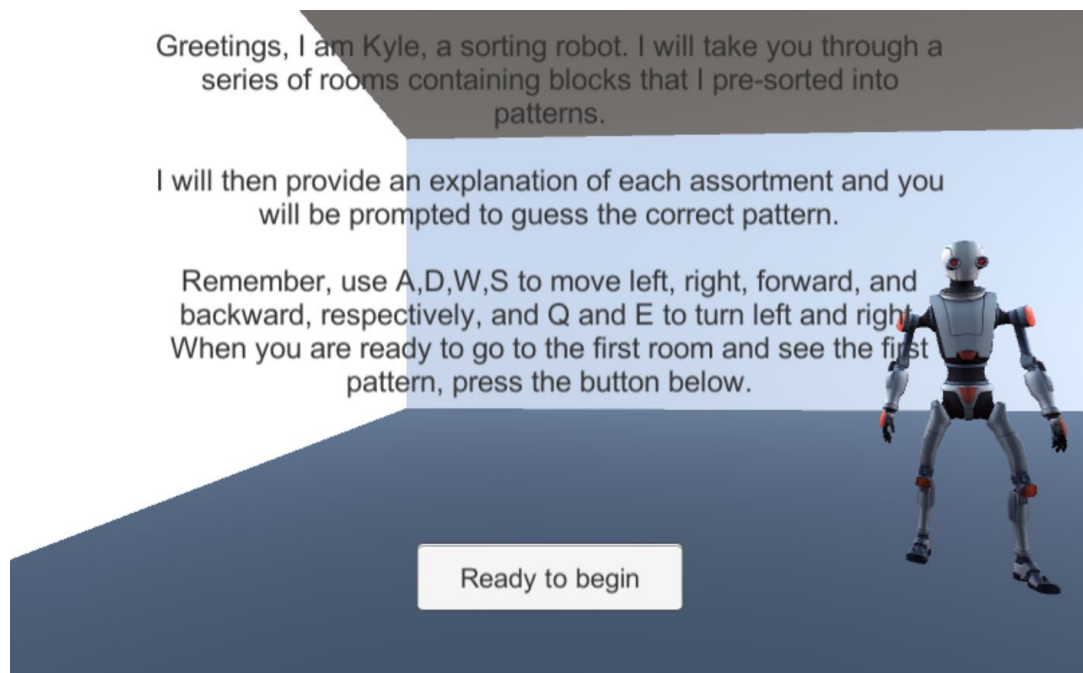


Figure 3. Introduction and Practice Room

The simulation began with an instruction screen (Figure 2) followed by the robot introducing itself and introducing the participant to the environment. The participant was then given an opportunity to practice navigating through the environment in a practice room with visible directions indicating how to use keys to move in the simulation (Figure 3). Once the participant was comfortable navigating within the simulation, the robot guided the participant to the first pattern in a new room. In every room, the robot presented the participant with a set of blocks that were sorted using a predetermined criteria. The robot then attempted to explain the criteria used to sort the pattern. The robot used a chat feature on the screen to communicate with the participant. The participant was then prompted to provide feedback in the form of a survey asking the participant to rate the quality of the robot's explanation using a five point Likert scale from very poor to excellent. Additionally, the participants were presented with a list of possible patterns to choose from and prompted to select the accurate pattern. The robot's explanation was provided to influence the participant's guess related to the correct pattern (Figure 4). This selection measured the participants' performance in pattern identification given the robot's explanations.

Explanation: "I am sorting by size and letter"

What is the pattern?

The pieces are sorted by color.
 The pieces are sorted by letter.
 The pieces are sorted by size.
 The pieces are sorted by color and size.
 The pieces are sorted by color and letter.
 The pieces are sorted by size and letter.
 The pieces are sorted in alphabetical order.

Rate the quality of the robot's explanation.

Excellent Good Fair Poor Very poor

Comments?

Enter text...

Submit

Figure 4. Survey Presented to Participants Following the Viewing of the Pattern and Robot's Explanation



Figure 5. Featured Pattern of Blocks Utilized in the Experiment, Sorted by Size

The robot guided participants through six different rooms with six different patterns created using blocks (Figure 5). Patterns were selected using blocks based on color, size, and letter, and could be either a single factor pattern, such as being sorted by color or in alphabetical order, or a two factor pattern, such as sorting by both color and letter. In order to determine

whether participants could distinguish between different qualities of explanation, a guideline was established for the creation of good, medium, and poor explanations. These guidelines were based on the idea that the more information the explanation provided, the better the explanation was. Using this logic, a good explanation provided complete information about the pattern; if the pattern was sorted by color and letter, the explanation would be, “I sorted by color and letter”. A medium level explanation provided some information about how the pattern was sorted, but not complete information. An explanation of this type either gave partial information about how the blocks were sorted by explaining one of the sorting factors but not both, or provided information about how the blocks were not sorted. In other words, a medium explanation for a pattern sorted by color and letter could be, “One of the attributes I sorted by was color”, or, “I did not sort by size”. Finally, a poor explanation provided no pertinent information about the pattern, but instead shared an irrelevant opinion or fact. For example, a poor explanation for a pattern sorted by letter would be, “Q is my least favorite letter”, or, “Q is the least used letter in the English alphabet”. This logic was applied to the six patterns used in the experiment, including two patterns for each level of explanation (see Table 5 in Discussion). The patterns and the level of explanation the robot used to explain the patterns were randomly selected. The order in which the patterns were presented were also randomly generated, and every participant encountered the same patterns and explanations in the same order.

This time, you will create a pattern and explain it to me.
Please use the blocks in this room to create your own pattern
using a minimum of 6 blocks and 2 piles.

Please drag and drop the blocks on the green platform. Do not
worry about lining the blocks up perfectly. To begin creating
your pattern, press the 'Start creating pattern' button.

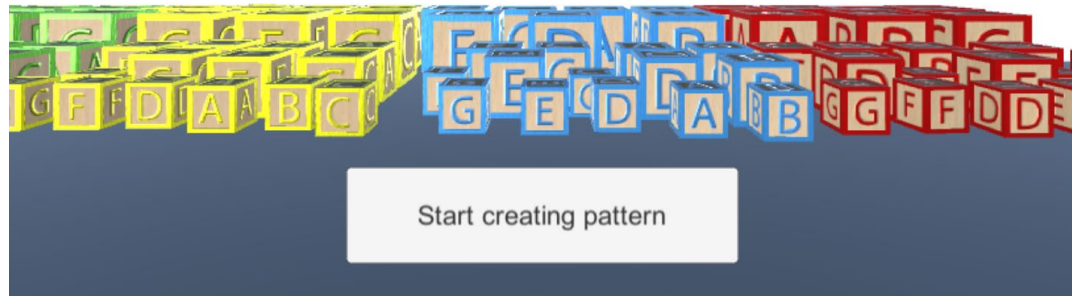


Figure 6. Participant View of Available Blocks for Pattern Creation

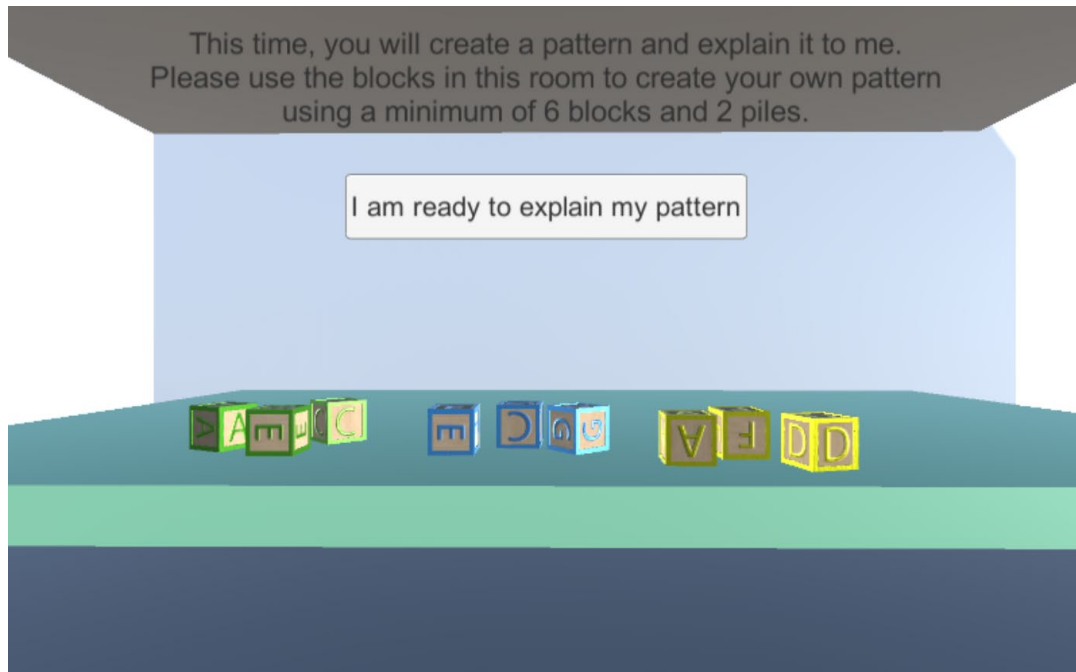
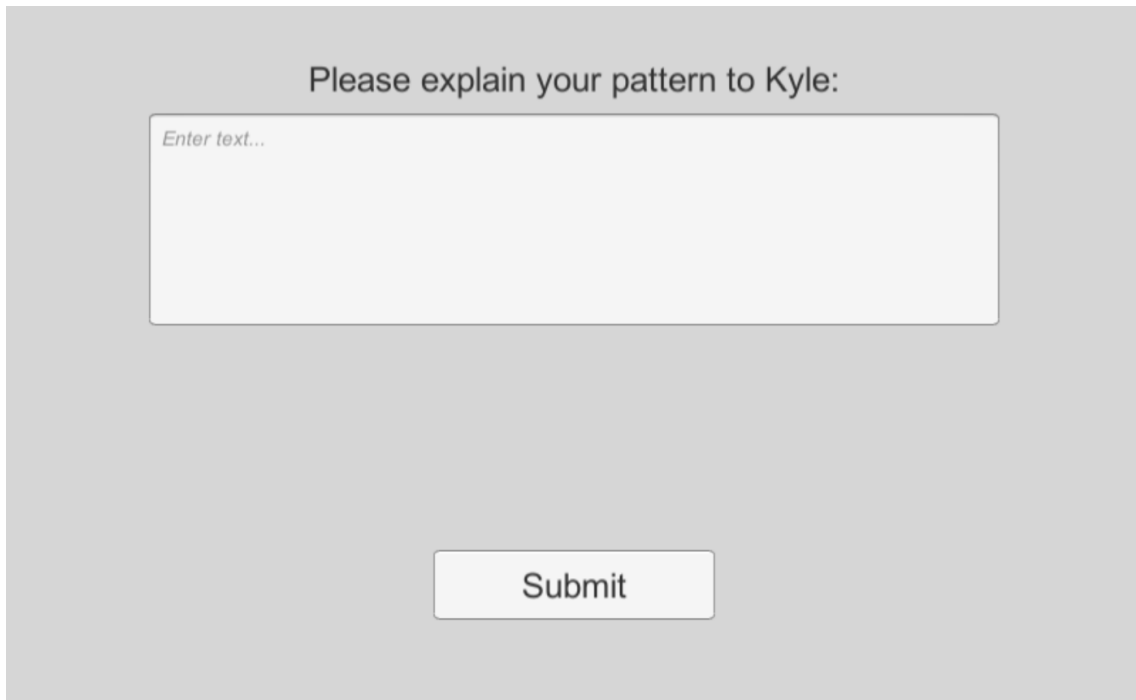


Figure 7. Example of a Drag-and-Drop Pattern



The image shows a chat interface on a light gray background. At the top, the text "Please explain your pattern to Kyle:" is centered. Below this is a large white text input box with a thin gray border and the placeholder text "Enter text...". At the bottom center of the interface is a white rectangular button with a thin gray border and the text "Submit" centered on it.

Figure 8. Chat Feature for Participants to Explain Pattern to the Robot

After the robot led the participants through the six rooms containing the presorted patterns, the robot led the participants into a final room with a bank of blocks of various sizes, colors, and letters. Participants were asked to create their own pattern using a drag and drop technique and then explain it to the robot using a chat feature (Figure 6, 7, 8). The data collected from this final room provided information about possibility of an explanatory convergence between the human participants and the robot. Participant-provided explanations were compared with the six robot-provided explanations and analyzed using the Linguistic Inquiry and Word Count software, developed by Pennebaker [5]. The software package analyzes text inputs and outputs word count, part of speech, categories of conversation touched upon, and four overall output variables that provide data on linguistic and communication style of the speaker. In this case, participant explanation inputs were run through the software and averaged, as were the robot's explanations. The four main variables produced by the Linguistic Inquiry and Word

Count software are analytical thinking, clout, authenticity, and emotional tone. Analytical thinking is a dimension that interprets the speaker's level of formality and logical thinking. Clout quantifies the confidence or leadership status of the speaker, and should not be confused with the power dynamic of the speaker in conversation. Authenticity reflects the speaker's tendency to communicate in a more humble, personal manner, and emotional tone conveys the speaker's positive or negative tone in communication. These four variables provide insight as to the participants' demeanor and attitude during the experiment and whether participants displayed any convergence or similarity in their method of generating explanations. Such an analysis provides a baseline for understanding how humans may converge with a robot's linguistic style and opens the door to future studies of linguistic accommodation.

Participant-provided explanations were analyzed and labeled as good, medium, or poor using the same criteria used for the robot's explanations. Participant's patterns were also divided into single factor and two factor patterns. Distinguishable patterns were identified according to the factors of the pattern and matched back to the explanations provided by the robot and the patterns the robot featured. Analyzing participants' explanations and chosen patterns allows for a better understanding of participant's comprehension, retention, and decision-making regarding how to best communicate a pattern to the robot so that the robot will understand the pattern. The experiment concluded with impressions of the robot using five point Likert scales on likeability and intelligence followed by demographic information.

Chapter 4

Results

Ratings of Robot-Provided Explanation

Participants followed the robot through each room and received six explanations from the robot describing six different patterns and rated the quality of the robot-provided explanations using a five point Likert scale. As time went on, participants were better at distinguishing between different levels of explanations. The results of the ratings of the six explanations are summarized in Table 2.

Table 2. Five Point Likert Scale Ratings of Explanations Presented by the Robot

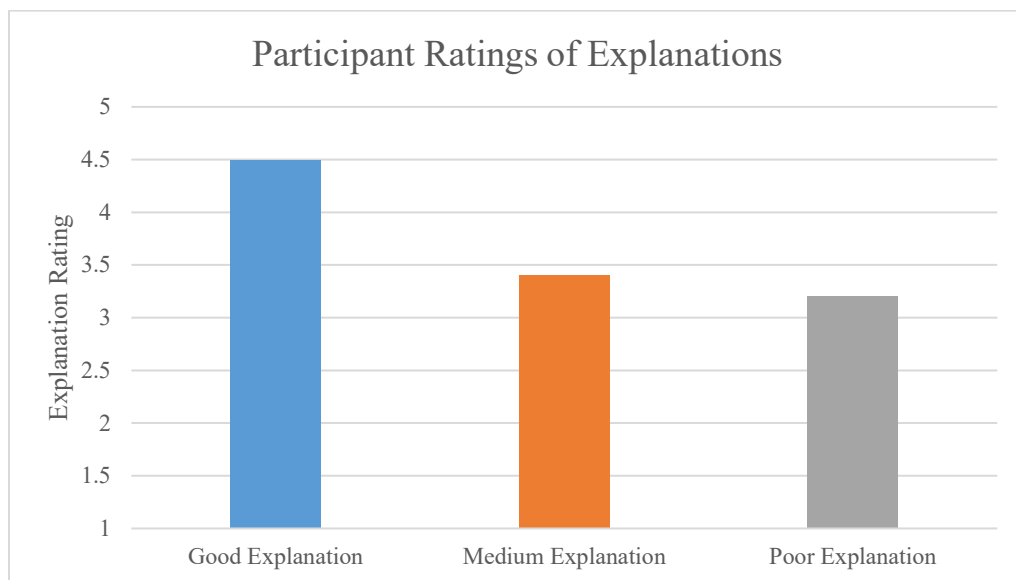
Explanation	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1	4.370	.186	3.988	4.753
2	3.444	.284	2.860	4.028
3	3.000	.239	2.509	3.491
4	3.815	.207	3.389	4.240
5	2.963	.331	2.282	3.644
6	4.630	.121	4.381	4.879

Grouping the explanation ratings for each level of explanation, there is a significant difference in means of good explanations ($M= 4.500$, $SD= .111$), medium explanations ($M= 3.407$, $SD= .166$), and poor explanations ($M= 3.204$, $SD= .219$), $F(1.652, 1.449)= 21.927$, $p<.001$, determined through a repeated measures ANOVA test. An ad hoc paired t-test applying the Bonferroni correction reveal that there is a significant mean difference between good and medium explanations, $t(27)= 6.076$, $p<.001$, and good and poor explanations $t(27)= 5.041$, $p<.001$, but not between medium and poor explanations, $t(27)= 1.097$, $p= 0.833$.

Table 3. Pairwise Comparisons of Levels of Explanation

(I) Explanation	(J) Explanation	Mean Difference (I – J)	t	Significance
1	2	1.093*	6.076	.000
	3	1.296*	5.041	.000
2	1	-1.093*	6.076	.000
	3	.204	1.097	.833
3	1	-1.296*	5.041	.000
	2	-.204	1.097	.833

1 = good explanation, 2 = medium explanation, 3 = poor explanation

**Figure 9. Participant Ratings of Explanations from 1 (very poor) to 5 (excellent)**

Pattern Identification

In addition to rating the robot's explanations, participants were asked to select the correct pattern based off of the robot's given explanation and the assortment of blocks presented. Such a selection provides a means of measuring the participants' performance in identifying the correct pattern given the robot's explanation. Unexpectedly, participants were able to most accurately select the correct pattern when the robot-provided explanation was poor and least accurately

select the correct pattern when the robot provided medium level explanations, $\chi^2 (2) = 29.400$, $p < .001$.

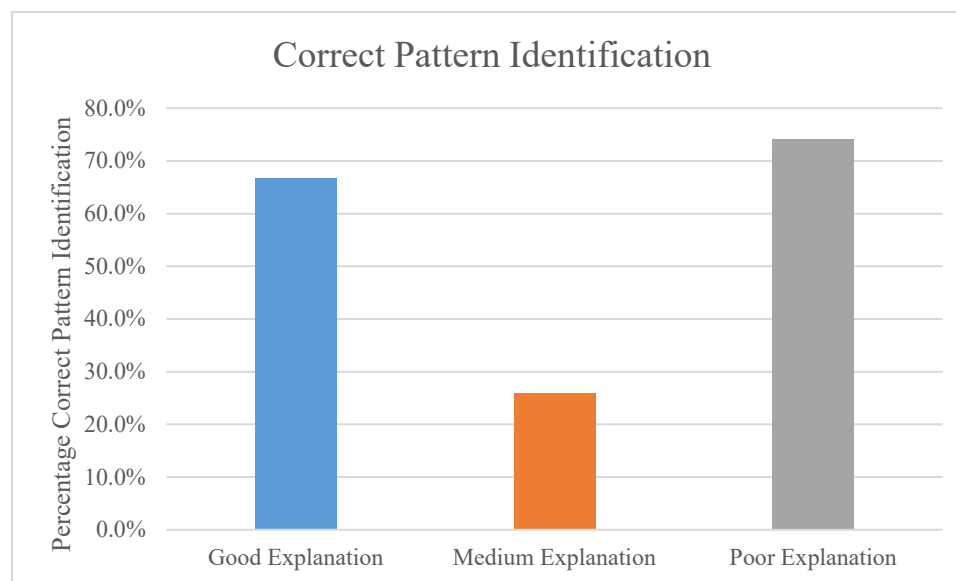


Figure 10. Correct Pattern Identification by Explanation Level

Throughout the experiment, the robot sorted four single factor and two two-factor patterns. A single factor pattern has only one component to it, whereas a two-factor pattern has two components to it that a participant must accurately identify. The number of factors in the pattern significantly impacted the pattern identification accuracy, $t(161) = 2.279$, $p = .024$. The findings are presented in Table 5.

Table 4. Participant Identification of Patterns

		No		Yes		Total	
		% within Type of		% within Type of		% within Type of	
		Count	Explanation	Count	Explanation	Count	Explanation
Type of Explanation	Good	18	33.3%	36	66.7%	54	100.0%
	Medium	30	74.1%	24	25.9%	54	100.0%
	Poor	14	25.9%	40	74.1%	54	100.0%
Total		72	38.3%	90	55.6%	162	100.0%

Communication and Linguistic Style

A total of 23 participants completed the pattern creation and participant-generated explanation portion of the experiment, with the average age between 25-34 and 87.0% male. Participant-provided explanations ($M= 7.22$) were approximately the equivalent length to the robot-provided explanations ($M= 7.50$), resulting in a percent difference of 3.80%. According to analytical thinking, participant explanations (74.63 out of 100) were rated as very high, and the robot-provided explanations were also rated as high (58.43 out of 100). Participant explanations (50.49 out of 100) were rated as relatively low in clout, while the robot-provided explanations were rated even lower in clout (28.00 out of 100). In terms of authenticity, participant explanations (29.60 out of 100) were rated as very low, as were the robot-provided explanations (19.80 out of 100). In emotional tone, participants received a score of 37.42, while the robot received a score of 50.18, neither ranking highly for a positive tone while providing explanations. Participant explanations (37.42 out of 100) were rated as low for emotional tone, while the robot-provided explanations (50.18 out of 100) were rated as low but higher than the participant explanations. Of the participants' explanations, 69.57% of the explanations were good, 21.74% were medium, and 8.70% were poor (Figure 11). 60.87% of the participant-provided explanations explained single factor patterns, 30.43% were sorted by color, 21.74% were sorted by letter, and 8.70% were sorted by size. 21.75% of the patterns were sorted using two factors, with 8.70% sorted by letter and size, 8.70% sorted by letter and color, and 4.35% sorted by color and size. Selected pattern-types are summarized in Figure 12.

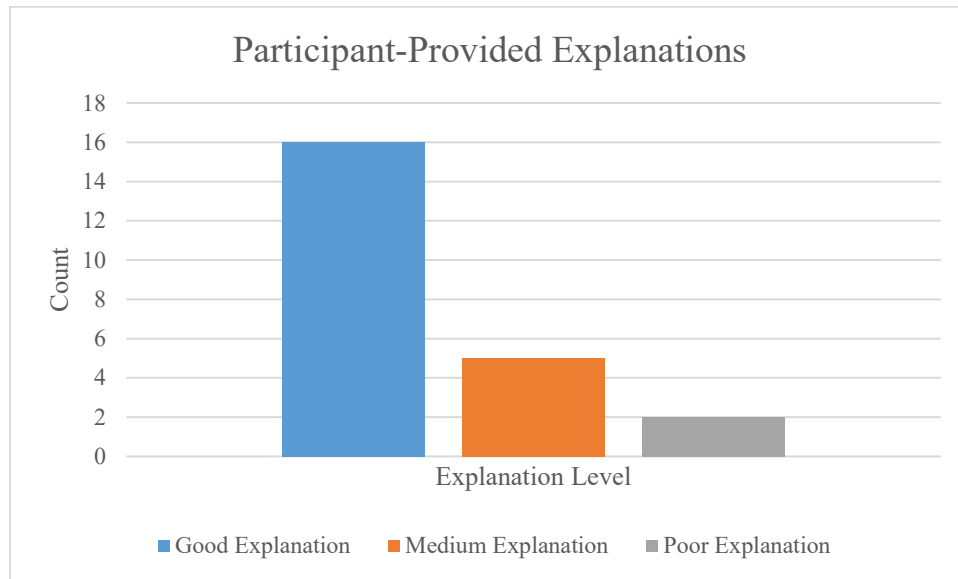


Figure 11. Participant Provided Explanations

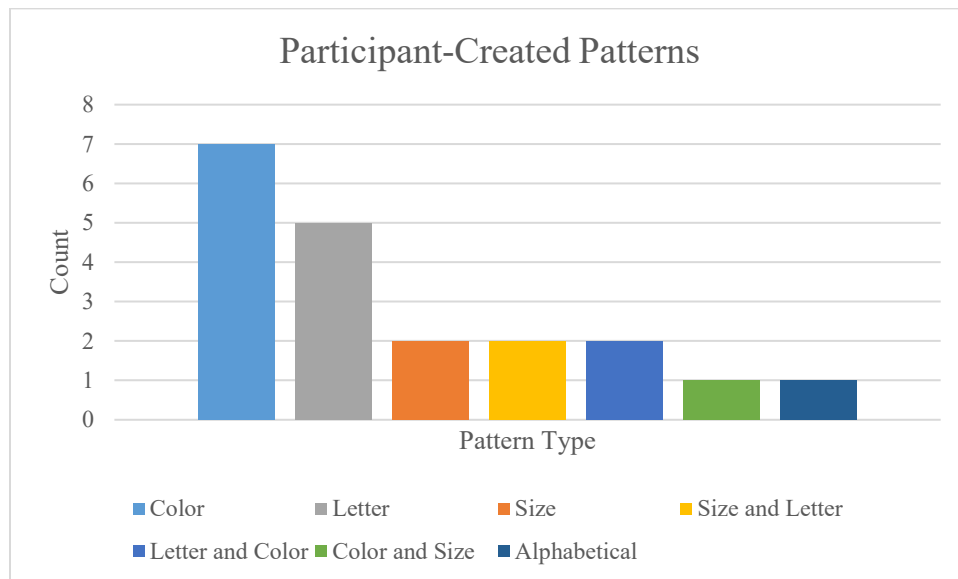


Figure 12. Participant-Created Patterns

Chapter 5

Discussion

Ratings of Robot-Provided Explanation

We found that participants rated the three levels of explanation differently. Good explanations were rated the highest (see Figure 9), followed by medium explanations, and poor explanations were rated the lowest. This supports H1, that the better the explanation, the higher the participants rated explanation provided by the robot. We also find that the levels of explanations match the ratings given by the participants, indicating that participants are in agreement with the level of explanations determined by the research team to be provided by the robot. Had the participants rated the three levels of explanations as of similar quality or of quality that did not reflect our predetermined levels, it would have been impossible to interpret the validity of the comprehension given a flawed rating of explanation quality. Furthermore, these results convey that participants were cognizant of quality of the explanations provided to them and could thus use the information contained in that explanation to identify the correct pattern.

Pattern Identification

We now explore how participants' used the explanations to evaluate the patterns. Our data here does not support H2, which stated that participants will be more likely to identify the correct pattern the better the explanation is. We find that participants better explanations do not

lead to more accurate pattern guessing. Surprisingly, poor explanations resulted in the most accurate pattern identification. A poor explanation contained no hints or direct information related to the actual pattern, and provided irrelevant information. We therefore examined how the number of factors contained in the patterns impacted the correct identification of the patterns when poor explanations were provided. We found that there was a significant difference between correct identification and number of factors in the pattern. It seems that the more factors contained within the pattern, the lower the pattern identification accuracy. As presented in Table 4, there were zero two-factor patterns in the poor explanation condition, aligning with this finding. The patterns were randomly generated and predetermined for the experiment. Hence, there was no assurance that each level of explanation would have the same number of single and two-factor patterns. The two poor explanations were both single factor patterns and, while they were poor explanations, were easy enough to determine the correct pattern despite the poor explanations provided by the robot. The good explanation and medium explanation conditions, however, had one single factor and one two-factor pattern, but the correct identification of the pattern was significantly higher for the good explanations. The good explanations condition provided explicit statements of the pattern containing complete information, while the medium patterns contained only partial information about the pattern. Despite having the same breakdown of single and two-factor patterns, the information in the good explanations condition allows for a person to more accurately identify the pattern than the partial information included in the medium explanations.

Table 5. Robot-Provided Explanations

Room	Explanation	Level of Explanation	Number of Factors in Pattern
1	I am sorting by size and letter.	Good	2
2	My favorite color is red.	Poor	1
3	I do not like to sort by color.	Medium	1
4	One of the attributes I am sorting by is size.	Medium	2
5	E is the most commonly used letter in the English alphabet.	Poor	1
6	I am sorting by size.	Good	1

Preferred explanation types are unique and frequently preference-based [7]. Providing any explanation at all as to how the patterns were sorted, whether grounded in factual statements or observations, builds common ground between the robot and participant and allows for increased trust and engagement with the robot [3], [13]. These results reveal that for explanations in which partial or incomplete information is given, a person's understanding of the pattern is lower. However, the inconsistency in number of single factor and two-factor patterns used for each explanation level is evidence of a flaw in the experiment and at this time a connection between explanation quality and participant performance cannot be determined. To amend this, the experiment should be rerun being sure to include one single-factor and one two-

factor pattern for each type of explanation. Doing so would allow for a more complete understanding of the relationship between explanation quality and performance of participants.

Communication and Linguistic Style

The final part of the experiment asked participants to generate explanations. This portion of the experiment was meant to provide insight into the possible matching of communication style between the robot and participants at the end of the experiment. The average word counts of explanations from both the participants and the robot were very similar, possibly conveying that participants felt they were able to adequately explain their pattern with the same number of words that the robot used. It may be the case that such a response indicates a convergence in communication style based on the type of information needed to be shared. In terms of analytical thinking, participants scored relatively similarly to the robot, potentially indicating that participants matched the robot in formality. By providing various explanations and giving a person the chance to familiarize themselves with the communication style of the robot, it may be the case that the person identified (presumably subconsciously) the robot's preferred style of communication and then applied the style to their communication with the robot when asked to generate an explanation of the robot.

Emotional tone scores for both the robot and the participants can be categorized as not outwardly positive, which was not surprising due to the fact that the communication between the robot and participant was an explanation about a pattern of blocks. Similarly to emotional tone, the robot and participants received a low score in authenticity. Neither were vulnerable nor revealed personal information during communication, likely once again reflecting the task of

explaining a pattern of blocks. In this case, vulnerability is not necessary to explain a pattern, but rather facts and informational statements should be used, aligning with the higher analytical thinking scores. Among all of the linguistic style indicators, clout was the category that resulted in the largest difference in scores between the robot and participants. Participants scored much higher in this category, suggesting that the person portrays more confidence when explaining. Additionally, this higher score may imply that the person prefers to take on a more dominant role when communicating with a robot.

Across linguistic categories, there was generally a level of matching in communication styles between the robot and participants. This supports H3A, as participants generally matched communication styles while explaining their pattern to the robot. However, convergence over time could not be determined due to the limited participant inputs and subsequent data available. Future work studying longer conversations and increased turn taking would more definitively determine convergence in a person's communication style based off the robot's speech. For instance, allowing the robot and the participant to take turns explaining patterns rather than having the robot present all of its patterns first would provide more data to find significant results.

The majority of participants explained their patterns using good explanations, indicating a preference for complete information when explaining to an unfamiliar robot. This finding supports H3B, meaning that participants preferred to explain their pattern in such a way that it matched the best explanation level the robot provided. Interestingly, more participants explained their patterns with a medium explanation than a poor explanation, which does not match the comprehension of participants during the pattern identification portion of the experiment. However, these results do agree with the participants' ratings of the robot-provided explanations.

This indicates that there may be inconsistencies in how a person prefers to give an explanation and how a person prefers to receive an explanation. Participants' pattern selection and explanation to the robot may reveal insight into their own preferences when conversing with a robot. Participants preferred to explain single-factor patterns, specifically patterns sorted by color and patterns sorted by letter, conveying a preference for explaining simpler patterns. The person may believe that communicating with a new individual, in this case a robot, warrants a simpler pattern with a more information-based explanation in order to facilitate the greatest understanding. Future work with rounds of turn-taking and explaining between the robot and participant may deliver intriguing results regarding the convergence between a person and robot during conversation. In addition, deeper insight into human preferences of communication styles when speaking with a robot may also be found in such studies.

Chapter 6

Conclusion

The intention of this experiment was to gain a better understanding of human performance given robot-provided explanations. The main findings from the experiment are that participants were able to consistently rate the quality of different levels of explanations. Additionally, participants had the highest performance in pattern identification accuracy when the explanation level was poor and the lowest pattern identification accuracy when the explanation level was medium. This can partially be explained by the number of factors in the pattern, either single-factor or two-factor, and by the amount of information provided in the explanations. Due to an inconsistency in single factor and two-factor patterns presented and explained by the robot, a connection between quality of explanation and participant performance cannot be determined.

Participants generally matched communication styles when providing their own explanations, but convergence over time could not be determined due to limited turn taking. In addition, participants preferred to explain their patterns using good explanations and single-factor patterns. As can be seen by these results, there may be inconsistencies in how a person prefers to give an explanation and how a person prefers to receive an explanation. The thought process and decision-making that goes into a person choosing how to sort a pattern and then how to explain it to the robot is beyond the scope of this experiment but still an important factor to consider in the future.

The results from this thesis presents important lessons learned for robot designers. When designing a robot with the ability to explain tasks or decision-making, explanations should contain full information and avoid instances where pertinent information is omitted, if possible. Additionally, simplifying the information that needs to be shared when possible can facilitate a higher level of performance.

The application of robots that can explain tasks and decision-making in a way that maximizes human performance is vast. For instance, the applicability of robots with this capability has the ability to transform the healthcare industry. Physician offices, outpatient centers, and hospitals can be streamlined through the utilization of task-based robots with explanation capabilities. Such robots could receive tasks from medical professionals, convey information to patients, refill prescriptions, or answer questions that patients have in a fashion that maximizes the chances of the patients' performance, which in this case may include following directions given by the robot such as taking a prescription at a certain time and following specific steps. Knowing how to best explain to humans to maximize performance expands the capabilities of robots can also be applied to future experiments in human-robot interaction were robot-provided explanation is necessary. Similarly, in the defense industry, it is critical that task-based information explained by a robot to a person is presented in a way that maximizes the person's performance of a certain order. Incomplete information or unclear explanations cannot be afforded in situations of national security or building and deployment of weapons systems. More generally, any application of robots with the ability to explain must aim for the highest human performance possible, as that will fulfill its purpose in task-based situations.

Future experiments can build upon the findings from this experiment. For instance, rerunning this experiment and including one single-factor pattern and one two-factor pattern for each level of explanation may gain a fuller understanding of which level of explanation leads to the highest performance in pattern identification among participants. Another possibility in terms of determining how to maximize human performance could be to present participants with a more challenging pattern or puzzle and have the robot explain in a series of randomized fashions, surveying participants after each explanation until the participant identifies the correct answer. In addition, this experiment has only just touched the surface of human-robot communication styles. Future projects could facilitate turn-taking between the participant and robot in creating and explaining patterns to determine if there is convergence over time in communication style. There is still a great deal to be uncovered in the study of the human dimension of robot-provided explanation in human-robot interaction, but these results provide a good basis to build upon.

BIBLIOGRAPHY

- [1] S. Nikolaidis, M. Kwon, J. Forlizzi, and S. Srinivasa, “Planning with Verbal Communication for Human-Robot Collaboration,” *ACM Trans. Human-Robot Interact.*, vol. 7, no. 3, pp. 1–21, 2018, doi: 10.1145/3203305.
- [2] A. Powers, A. D. I. Kramer, S. Lim, J. Kuo, S. L. Lee, and S. Kiesler, “Eliciting information from people with a gendered humanoid robot,” *Proc. - IEEE Int. Work. Robot Hum. Interact. Commun.*, vol. 2005, pp. 158–163, 2005, doi: 10.1109/ROMAN.2005.1513773.
- [3] N. Wang, D. V. Pynadath, and S. G. Hill, “Trust calibration within a human-robot team: Comparing automatically generated explanations,” *ACM/IEEE Int. Conf. Human-Robot Interact.*, vol. 2016-April, pp. 109–116, 2016, doi: 10.1109/HRI.2016.7451741.
- [4] J. Wang, A. Chellali, and C. G. L. Cao, “Haptic Communication in Collaborative Virtual Environments,” *Hum. Factors*, vol. 58, no. 3, pp. 496–508, 2016, doi: 10.1177/0018720815618808.
- [5] K. G. Niederhoffer and J. W. Pennebaker, “LINGUISTIC STYLE MATCHING IN SOCIAL INTERACTION,” doi: 10.1177/026192702237953.
- [6] M. A. Riordan, K. M. Markman, and C. O. Stewart, “Communication Accommodation in Instant Messaging: An Examination of Temporal Convergence,” *J. Lang. Soc. Psychol.*, vol. 32, no. 1, pp. 84–95, 2012, doi: 10.1177/0261927X12462695.
- [7] F. C. Keil, “Explanation and Understanding,” doi: 10.1146/annurev.psych.57.102904.190100.
- [8] A. Yamazaki, K. Yamazaki, Y. Kuno, M. Burdelski, M. Kawashima, and H. Kuzuoka,

- “Precision timing in human-robot interaction: Coordination of head movement and utterance,” *Conf. Hum. Factors Comput. Syst. - Proc.*, pp. 131–139, 2008, doi: 10.1145/1357054.1357077.
- [9] C. L. Sidner, C. Lee, L. P. Morency, and C. Forlines, “The effect of head-nod recognition in human-robot conversation,” *HRI 2006 Proc. 2006 ACM Conf. Human-Robot Interact.*, vol. 2006, pp. 290–296, 2006, doi: 10.1145/1121241.1121291.
- [10] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita, “Footing in human-robot conversations,” *2009 4th ACM/IEEE Int. Conf. Human-Robot Interact.*, vol. 2, no. 1, p. 61, 2009, doi: 10.1145/1514095.1514109.
- [11] M. M. A. De Graaf and B. F. Malle, “How people explain action (and autonomous intelligent systems should too),” *AAAI Fall Symp. - Tech. Rep.*, vol. FS-17-01-, pp. 19–26, 2017.
- [12] T. Nomura, T. Kanda, T. Suzuki, and K. Kato, “Prediction of human behavior in human - Robot interaction using psychological scales for anxiety and negative attitudes toward robots,” *IEEE Trans. Robot.*, vol. 24, no. 2, pp. 442–451, 2008, doi: 10.1109/TRO.2007.914004.
- [13] S. Kiesler, “Fostering common ground in human-robot interaction,” *Proc. - IEEE Int. Work. Robot Hum. Interact. Commun.*, vol. 2005, pp. 729–734, 2005, doi: 10.1109/ROMAN.2005.1513866.
- [14] M. Harbers, K. Van Den Bosch, and J. J. C. Meyer, “A study into preferred explanations of virtual agent behavior,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5773 LNAI, pp. 132–145, 2009, doi: 10.1007/978-3-642-04380-2_17.

- [15] A. R. Wagner, P. Robinette, and A. Howard, “Modeling the Human-Robot Trust Phenomenon: A Conceptual Framework based on Risk,” 2018, doi: 10.1145/3152890.
- [16] M. Hanheide *et al.*, “Robot task planning and explanation in open and uncertain worlds,” *Artif. Intell.*, vol. 247, pp. 119–150, Jun. 2017, doi: 10.1016/j.artint.2015.08.008.

ACADEMIC VITA

Sydney Hannah

EDUCATION

The Pennsylvania State University Schreyer Honors College	University Park, PA
Bachelor of Science in Mechanical Engineering	Graduation: May 2021
Bachelor of Arts in International Politics	
London School of Economics and Political Science- <i>Decline of the West in a New Asian Century</i>	<i>June - July 2018</i>
National University of Singapore- <i>Fundamentals of Product Development & Design</i>	<i>May - June 2018</i>

PROFESSIONAL EXPERIENCE

Northrop Grumman: Pathways Program Associate Mechanical Engineer	<i>September 2021</i>
ExxonMobil: Pipelines and Risers Engineering Intern	<i>June – Aug 2020</i>
<ul style="list-style-type: none">Validated internal riser and flowline engineering tools, one with 223 parameters and one with 668 parameters, and developed recommendations for enhancements for tool performanceCollaborated across teams to confirm utility of and feasibility for use of tools in generating internal front-end engineering design deliverablesOptimized internal expertise and saved 240+ hours of work on projects through implementation of tool	
ExxonMobil: Socioeconomic Management Engineering Intern	<i>May – Aug 2019</i>
<ul style="list-style-type: none">Created and launched a community of practice for operations in 25 countries to connect 182 practitioners in order to more effectively assess cultural, economic, and environmental impacts on communitiesLaunched a sustainable platform that expanded membership by over 400% within a year and expanded to host training sessions and spotlight best practices and lessons learned in socioeconomic managementCollaborated with practitioners to identify and mitigate negative impacts in communities near operations globally	
Research Assistant	
<ul style="list-style-type: none">Robot Ethics and Aerial Vehicles Lab<ul style="list-style-type: none">Analyze the ability of robots to explain tasks and actions comprehensibly to humans in order to expand the utility of robots in the workforce and societyInformation Sciences and Technology Lab<ul style="list-style-type: none">Examined information communication technology usage among Rwandan refugees using statistical analyses in the R programming language to link technology sharing, community, and self-efficacy	<i>Jan 2020 - present</i> <i>Feb 2018 – Dec 2019</i>

LEADERSHIP & INVOLVEMENT

Society of Women Engineers	<i>Aug 2018 - present</i>
<ul style="list-style-type: none">Officer of Corporate Engagement<ul style="list-style-type: none">Establish relationships between over 50 corporate contacts and over 250 active membersSpearhead, plan, and organize volunteers for the largest engineering career fair on the east coastCollaborate with officers to develop an adaptive, virtual model for SWE during an unprecedented semesterDirector of Professional Development<ul style="list-style-type: none">Propelled the careers of women in engineering by preparing and running events to develop professional skills such as resume building and career planningCreated and ran 3 new corporate events that contributed to chapter winning SWE award at the world's largest conference for women in engineering	<i>April 2020 - present</i> <i>April 2019 – April 2020</i>
Penn State Women in Engineering Program	<i>Aug 2017 - present</i>
<ul style="list-style-type: none">Networking Lead<ul style="list-style-type: none">Design and run an industry-sponsored virtual networking reception and career panel for 282 alumni, industry professionals, and women engineersMentor<ul style="list-style-type: none">Familiarized 180 incoming engineers with critical skills to excel in the academic and professional setting through resume-construction, networking, and engineering designMentored 15 engineers throughout the school year to facilitate adjustment to college	<i>March 2020 - present</i> <i>Aug 2019 - present</i>
Schreyer Consulting Group	<i>Aug 2017 - present</i>
<ul style="list-style-type: none">President<ul style="list-style-type: none">Execute opportunities for members to gain experience in approaching complex problem solving in case preparation with real-time feedback from partners in multinational firmsExpand reach of club from firms located within Pennsylvania to New York and Colorado	<i>Jan 2020 – Dec 2021</i>

HONORS & INTERESTS

- Honors:** Phi Beta Kappa Academic Honor Society, Engineering Underrepresented Fund (2019,2020), Louis A. Harding Memorial Scholarship (2019), President's Freshman Award (2018), Academic Excellence Scholarship (2017-2020), Dean's List (2017-2020)