AIDA: An Assistant for Workers with Intellectual and Developmental Disabilities

By

Ronald Moore

Submitted to the graduate degree program in the Department of Electrical Engineering and Computer Science and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Master of Science.

Chair: Andrew Williams

Member: Arvin Agah

Member: Michael Branicky

Member: Richard Wang

Date Defended: June 12, 2020

The thesis committee for Ronald Moore certifies that this is the approved version of the following thesis:

AIDA: An Assistant for Workers with Intellectual and Developmental Disabilities

Chair: Andrew Williams

Date Approved: June 12, 2020

Abstract

Roughly 1 in 5 people in the United States have an intellectual or developmental disability (IDD), which is a substantial amount of the population. In the realm of human-robot interaction, there have been many attempts to help these individuals lead more productive and independent lives. However, many of these solutions focus on helping individuals with IDD develop social skills. For the solutions that do focus on helping people with IDD increase their work productivity, many of these involve giving the user control over a robot that augments the worker's capabilities. In this thesis, it is posited that an autonomous agent could effectively assist workers with IDD, thereby increasing their productivity. The artificially intelligent disability assistant (AIDA) is an autonomous agent that uses social scaffolding techniques to assist workers with IDD. Before designing the system, data was gathered by observing workers with IDD perform tasks in a light manufacturing facility.

To test the hypothesis, an initial Wizard-of-Oz (WoZ) experiment was conducted where subjects had to assemble a box using only either their dominant or non-dominant hand. During the experiment, subjects could ask the robot for assistance, but a human operator controlled whether the robot provided a response. After the experiment, subjects were required to complete a feedback survey. Additionally, this feedback was used to refine and build the autonomous system for AIDA.

The autonomous system is composed of data collection and processing modules, a scaffolding algorithm module, and robot action output modules. This system was tested in a

simulated experiment using video recordings from the initial experiment. The results of the simulated experiment provide support for the hypothesis that an autonomous agent using social scaffolding techniques can increase the productivity of workers with IDD. In the future, it is desired to test the current system in a real-time human-subjects experiment before using it to assist workers with IDD.

Acknowledgements

I would like to thank my parents Ronald and Rhonda Moore for their continued support through my Master's program and my academic career as a whole. I would also like to thank my advisor, Dr. Andrew Williams, for his mentorship and encouragement and for showing me how to be an effective researcher. Additionally, I would like to thank Dean Arvin Agah, Dr. Michael Branicky, and Dr. Richard Wang for their invaluable feedback and participation on my thesis committee.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Intellectual and developmental disabilities are lifelong conditions that form during the developmental years and are denoted by below-average intellectual functioning and limitations in adaptive functioning [McKenzie et al.2016]. Intellectual functioning is defined as anything related to judgment, learning, problem solving, planning, and abstract thinking. Adaptive functioning relates to the abilities to conform to developmental and societal standards for independence and meet their social obligations [APA2015]. Adaptive functioning and the onset of these in early childhood are the major qualities of IDD defined by the American Psychiatric Association (APA) in the 5th edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-V) [Association. and Association.2013].

According to the DSM-V, the four levels of severity for IDD are mild, moderate, severe, and profound. Mild levels of IDD include difficulties with reading, writing, math, money, and time. Additionally, persons with mild IDD have problems with abstract thinking and short-term memory. These individuals are socially immature and have a limited understanding of risks in social situations. Individuals with mild IDD also need help with more

complex daily activities, including grocery shopping, transportation, food preparation, and money management. For individuals with moderate IDD, academic skill remains at an elementary level throughout adulthood. Additionally, they have difficulties perceiving social cues. Although they can perform basic tasks for themselves such as eating dressing and hygiene, they require additional teaching and time in order to do so successfully. Persons with severe IDD have little understanding of writing, math, money, or time. They generally need a caretaker to help them with problem solving. These individuals have a limited vocabulary and grammar and use single words and phrases while speaking. They require a caretaker, as they require assistance in daily activities, as they are incapable of making decisions related to the well-being of themselves or others. For individuals with profound IDD, conceptual skills involve the physical world rather than abstractions, and physical objects are often used in a goal-directed fashion for recreation, work, and self-care. They have a very limited understanding of speech or gestures and use non-verbal, non-symbolic actions to communicate their intentions. They are extremely dependent upon others for assistance with everyday activities, although they may be able to perform simpler tasks by themselves [Association. and Association.2013].

There are many people with IDD in the United States. According to a 2012 report conducted by Matthew Brault, there are roughly 15.2 million adults with some sort of cognitive, mental, or emotional disability [Brault2012]. Roughly 10.6 million adults had a mental or cognitive disability, including a learning disability, Alzheimer's disease, senility, or dementia. Roughly 1.2 million adults had an intellectual disability, and 944,000 had other developmental disabilities. The remaining 4.7 million adults had some other mental or emotional condition.

In regards to employment, roughly 51.9 percent of people with IDD have a job [Brault2012]. The median monthly income of these individuals was $1,619. Similarly, a 2020 survey conducted by Hussain Almalky found that roughly two-thirds (67%) of young adults with IDD

were employed, with the average wage being $10.40 per hour [Almalky2020]. Additionally, Almalky found that people with IDD that worked in sheltered employment earned significantly less weekly than those that worked in supported employment, with sheltered employment defined as work and support in a segregated setting and supported employment defined as work in an integrated setting with ongoing support services for workers with IDD.

1.2 Motivation

It can be seen that there is support for robotic assistance for individuals with IDD. Many of these individuals are without jobs either because they are incapable of learning a certain skill or do not have access to the extensive training they require. Additionally, the varying levels of IDD require caretakers to give different amounts of assistance, which may be an issue if a caretaker works well with individuals with a certain level of IDD. A humanoid robot could provide caretakers with an extra helping hand.

1.3 Thesis Outline

The upcoming sections of the thesis are ordered as follows. The next section in this work will be the *Related Work* section, which will provide an explanation for all the domains related to this research and examples of other research in those domains. These domains include human-robot interaction, social robotics, assistive robotics, social scaffolding, and convolutional neural networks. Succeeding the *Related Work* section will be the *Project Resources* section, which will provide detail about the robot, software tools, and machine learning frameworks used in this research. After the *Project Resources* section will be the *Approach* section, which will discuss the designs of the human-subjects and simulated experiments. This section will also provide details about the structure of the final AIDA system, including

the data collection and processing module, the social scaffolding algorithm, and the output action module. Following the *Approach* section will be the *Results* section, which will discuss the results of the human-subjects and simulated experiments. Next will be the *Discussion* section, where the results of the human-subjects and simulated experiments will be analyzed. The final sections are the *Conclusion* and *Future Work* sections. The *Conclusion* section will summarize the work presented in this paper, while the *Future Work* section will talk about the potential future directions for the AIDA system.

Chapter 2

Related Work

2.1 Human-Robot Interaction

Simply put, human-robot interaction (HRI) is the interaction between humans and robots. Some popular subsets of HRI are social robotics, assistive robotics, and collaborative robotics.

2.2 Social Robotics

Social robotics involves autonomous robots interacting with humans and other autonomous robots by adhering to social behaviors related to their tasks. This includes the ability to have verbal conversations with humans and to understand human emotions and to express their own emotions in a verbal and non-verbal manner.

Researchers have looked at increasing the social capabilities of robots in order to provide a more enriched interaction with humans. Michael de Jong and his team proposed a framework to increase the social learning capabilities of an autonomous agent [de Jong et al.2018]. The system focused on vision, speech, and other modalities. For vision, the authors used the OpenPose human pose recognition and the You Only Look Once (YOLO) and Faster R-

CNN deep learning models for robust object recognition. Regarding speech, they used the robot's built-in speech recognition system along with Google Cloud Speech, a cloud-based speech recognition software, and external microphones to further increase the accuracy of the robot's speech recognition. The authors also incorporated the robot's tablet to allow for tactile inputs from a human user. The authors also created a mobile application to allow the human user to give smartphone commands to the robot. Despite this profound research, they did not work to enhance the productivity of workers with IDD. On the contrary, it aimed to help fully-capable people with everyday tasks.

2.3 Assistive Robotics

Assistive robotics observes how robots can help the disabled and elderly perform daily tasks. There has been extensive research conducted in the domain of assistive robotics. Siddarth Jain and Brenna Argall constructed a probabilistic model that allows a teleoperated robotic arm to infer the most likely intent of the human user [Jain and Argall2019]. Unlike the research in this thesis, research conducted by Jain and Argall did not use an autonomous humanoid robot and was not intended to help individuals with IDD in the workplace.

Similarly to Jain and Argall, Laura Herlant and her team built a model for a teleoperated robotic arm that allows the arm to automatically switch modes by inferring the intent of the human user [Herlant et al.2016]. The research conducted by Herlant and her team did not use an autonomous humanoid robot, nor is it intended to be used by people with IDD in the workplace.

Edmanuel Cruz and his group developed a system to help the elderly and intellectually challenged complete their daily tasks, resulting in increased levels of independence and quality of life for the individual [Cruz et al.2018]. For a given scheduled task, the robot gives the user the option to complete the task now or later. If the user chooses to complete the

task now, the robot leads the user to the location where the task should be performed before giving instructions on how to complete the task. Although this research aimed at helping people with IDD, this work did not focus on helping people with IDD increase their work productivity.

Kazuaki Takeuchi and his group devised a solution in which a person with a disability was given control over a robot avatar that was able to perform tasks [Takeuchi et al.2020]. Depending on the disability, users can control the robot through a graphical user interface or through using their gaze. However, this research differs from the work of Takeuchi et. al in that it offers a solution where the robot is autonomous and offers help to an individual with IDD as opposed to those with only physical disabilities.

2.4 Social Scaffolding

Regarding scaffolding, there have been extensive studies on how this dynamic occurs between humans. Michael Mascolo [Mascolo2005] introduced a new perspective on the scaffolding model used to explain how an adult expert helps a child learn a new skill. The traditional definition of scaffolding focuses its attention on the structuring actions that the expert gives to the child. Mascolo's model of scaffolding defined it as a dynamic coactive process between a child and an expert adult. Social scaffolding observes how exchanges with other people guide development in new directions. Mascolo concluded by stating that in order for his model of scaffolding to be effective, moment-by-moment observations must be made of coactions between a person and their environment. Mascolo's research focused on the scaffolding dynamic between humans. Similarly, when Sonia Chernova and Andrea Thomaz discussed teaching between robots and humans, they mention dynamic scaffolding, which involves the teacher adjusting the level of information they give to the learner [Chernova and Thomaz2014]. This research focuses on adjusting the level of assistance that

the robot provides to an individual with IDD.

There have been applications of social scaffolding through collaborative tasks with robots. Cynthia Breazeal and Andrea Thomaz introduced a learning mechanism called socially guided exploration that allows an agent to learn on its own using a reinforcement learning algorithm and learn from the social scaffolding of a human teacher [Breazeal and Thomaz2008]. The main behaviors for the system are novelty, mastery, and exploration. The novelty behavior is triggered when the agent is introduced by a human teacher to a new task that is not contained in its task set. The mastery behavior is activated when an agent's ability to perform a specific task is low. The exploration behavior is started when the agent is trying to figure out whether a task will achieve a certain goal. Results from the experiments showed that the robot learns significantly faster when a human teacher is involved than when a robot is trying to learn on its own. This research investigated scaffolding from the perspective of the human as the teacher and the robot as the learner. This research, however, looks at scaffolding with the robot as the teacher and the human as the student.

Frederic Kaplan and his team investigated scaffolding but using clicker training to teach an animal-like robot complex behaviors [Kaplan2001]. The first step of clicker training is to get the animal familiar with secondary positive reinforcement. Next, the animal is guided slowly through the desired action. After this has been done, the trainer will associate the action with a command word. Finally, the robot will attempt to perform complex behavior by itself. The trainer will only provide secondary positive reinforcement when the animal performs the exact behavior. In the experiments, participants were able to successfully teach the robot dog new tricks. Similar to [Breazeal and Thomaz2008], the work by Kaplan et. al involved the human as the teacher and the robot as the student, while this work involves a robot teacher and a human student.

Andrea Thomaz and Cynthia Breazeal created a new model for autonomous agent learning [Thomaz and Breazeal2008]. While the agent primarily uses reinforcement learning tech-

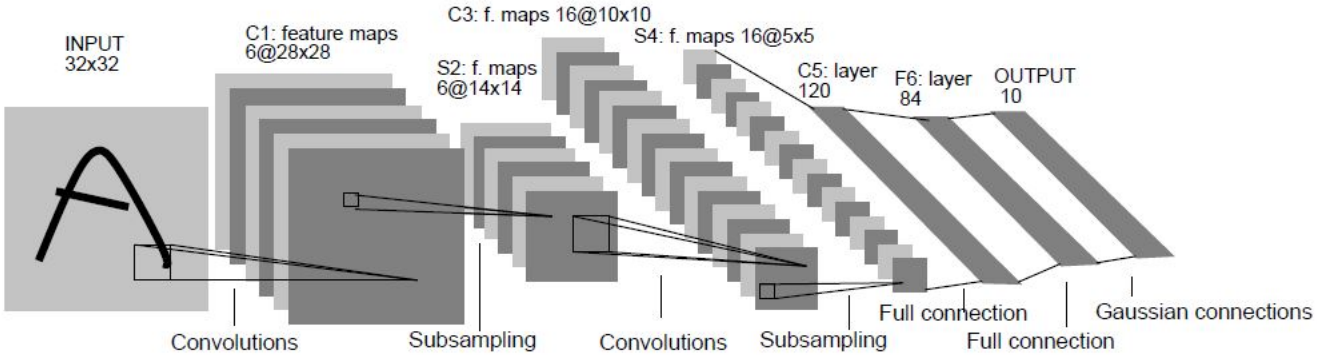

Fig. 1. Architecture of LeNet-5, a Convolutional Neural Network, here for digits recognition. Each plane is a feature map, i.e. a set of units whose weights are constrained to be identical.

Figure 2.1: CNN architecture from [LeCun et al.1999]

niques to learn, it can also utilize social learning cues given by human teachers to learn faster. From the experimental results of the initial model, the authors made a few key observations. First, they found that human teachers wanted to draw the agent's attention to certain things in a task in order to guide the agent's exploration. Next, they observed that people had a positive bias in their rewarding behavior, suggesting that they wanted to motivate the agent during their learning endeavors. Another important observation made by the authors is that human teachers benefited from learners' feedback in the form of social cues. Finally, the authors observed that when human teachers believed their negative feedback was being used by the learner, then they continued to give this negative feedback when appropriate. These observations were then tested in further experiments, and the results showed that these observations from the initial experiments held true. The work presented by Thomaz and Breazeal focused on a human teaching a robot, whereas this work focuses on a robot teaching a human.

2.5 Convolutional Neural Networks

The neural network used to detect emotion in a human is a convolutional neural network (CNN). A CNN is a special type of neural network that is commonly used working with enriched data such as images, videos, and audio samples. Early research on CNNs was conducted by Kunihiko Fukushima [Fukushima1980], but the findings of Yann LeCun and his team at ATT Shannon Lab allowed for gradient-based learning for CNNs [LeCun et al.1999]. The basic architecture can be viewed in 2.1. Each unit in the layer of the CNN receives inputs from units located in a small neighborhood of the previous layer. The units in a layer are placed into planes. Units within the same plane share the same weights, and the outputs of these units are called feature maps. The feature maps gather pertinent information about the features of a given image. LeCun found that the feature of the image is significantly more useful than its exact location in the image, and therefore searched for a way to lower the spatial resolution of the feature map to discard the information of the feature's exact location in an image. He achieved this by using sub-sampling layers that use local averaging and sub-sampling on the feature maps in the previous hidden layer of the CNN. The residual network (ResNet) architecture will be the CNN architecture used in this research [He et al.2015].

Chapter 3

Project Resources

3.1 Pepper Robot

The Pepper robot is a humanoid robot developed by Softbank Robotics that is capable of having basic interactions with humans. Pepper possesses 20 degrees of freedom to enable it to perform human-like actions. It also possesses both 2D and 3D cameras that allow it to process images and videos as well as perform autonomous navigation. Infrared sensors that add extra support for autonomous navigation. It also has built-in text-to-speech and speech recognition modules, allowing it to speak up 15 different languages and to recognize these languages. It also contains touch sensors, LEDs, and microphones. The touch sensors allow for a more sociable interaction with humans. The LEDs give the robot the ability to convey different intentions and emotions in a nonverbal manner. The microphones allow for human communication with the robot and the possibility for voice-command integration. More details on the Pepper robot can be found in [Pandey and Gelin2018].

3.2 IBM Watson Speech-to-Text

The IBM Watson Speech-to-Text is a product provided by IBM which is capable of translating audio into text. Speech recognition is a common application of speech-to-text that allows a robot to have a social conversation with a human user. According to the research done by Thomas et. al, IBM's speech recognition model is comprised of both an acoustic model for the audio input and a language model for natural language processing (NLP) [Thomas et al.2019]. The acoustic model is composed of both a residual network (ResNet) and a long short-term memory (LSTM) network. The outputs of these two neural networks are then combined to form the final decoding of the audio input. The language model uses a hybrid of an n-gram model and feed-forward neural network (NN) for initial decoding and an LSTM model for rescoring the decoded output of the n-gram-feed-forward NN model.

IBM's speech recognition engine is being used in place of the one provided by Softbank Robotics due to its simplicity and robustness. Pepper's built-in speech recognition engine requires the user to manually create the dictionary of words that the robot should recognize. In contrast, IBM's speech recognition has been pre-trained to accurately recognize over 80,000 words in the English dictionary. It is capable of performing both real-time transcription and transcription on audio files. For the purposes of this research, the transcription of audio files was used.

3.3 IBM Watson Tone Analyzer

The IBM Watson Tone Analyzer is a service created by IBM that provides detailed sentiment analysis for a given section of text. According to research conducted by Yichen Wang and Aditya Pal, the Tone Analyzer uses an inference algorithm to find the most probable sentiment expressed by the input text [Wang and Pal2015]. The algorithm is optimized by the added constraints of multiple emotions, topics, emotion bindings, emotion lexicon and

ordering, and bias.

The three major tones that the service provides analysis for are the emotional, social, and language tones. The emotional tone tries to determine the prevalence of the five major emotions in the given text. These emotions are anger, disgust, fear, joy, and sadness. The social tone looks at the openness, conscientiousness, extroversion, agreeableness, and emotional range being expressed in the text. The language tone examines the underlying language style of the text and determines the analytical, confident, and tentative nature of the style. For the purposes of this research, only the emotional tone results will be used.

3.4 PyTorch

This research uses PyTorch, an open-source deep-learning framework created by Facebook's AI Research Lab [Paszke et al.2019]. Their main goal is to achieve dynamic eager execution without sacrificing performance. When designing PyTorch, the researchers wanted to make the library as Pythonic as possible, considering that Python is the most popular language used in the deep learning communities. Additionally, they aimed to make the library easy for researchers to use, taking away much of the complexity of defining various deep-learning components. This makes PyTorch the optimal choice for research projects.

Another popular open-source deep-learning framework is Tensorflow, which was created by AI researchers at Google [Abadi et al.2016]. One of the design principles used for Tensorflow was to construct dataflow graphs from primitive operators. This was to make it easier for users to construct graphs and custom optimization algorithms. Another design principle used was deferred execution, which requires the entire graph of the model to be defined before it can be trained and optimized. Although this ensures higher GPU utilization, it also makes it harder for users to define more complex models such as recurrent neural networks (RNN). Tensorflow is a more ideal choice for large-scale machine learning applications.

With PyTorch, developers can create their own custom frameworks or use robust pre-trained models such as AlexNet, ResNet, InceptionV3, and other networks. For the purposes of this research, a pre-trained model will be used, as this will make it easier for the model to achieve reasonable accuracy.

Chapter 4

Approach

4.1 Data Collection

For data collection, the Cottonwood, Inc. manufacturing facility in Lawrence, Kansas was visited. Cottonwood employs people with IDD. During the visit, after receiving their legal consent, workers were recorded performing their work tasks. Some workers placed labels on canned pet food, some constructed belt fasteners, while others assembled cardboard boxes. Additionally, 10-minute interviews were conducted with a few of the workers. The workers were asked general questions about their family, how long they had been working at Cottonwood. In addition to these questions, workers were also asked what they felt could make their job easier. More details can be found in [Williams et al.2019].

4.2 Observations

While observing the workers in the facility, it was noticed that many of the workers performed their tasks at work stations with groups of other workers. Another key observation made during the visit was that workers' levels of IDD varied in the facility. Workers with severe or

Figure 4.1: Subject interacting with AIDA during experiment

profound levels of IDD needed staff members to assist them in every part of a task. Workers with moderate levels of IDD only needed to assist them by providing words of encouragement to help the workers complete a task or regain focus on the task. In other parts of the facility, staff members stood back while workers with mild levels of IDD operated heavy machinery independently.

In regards to the interview, many of the workers said they would not want someone doing their work for them and they would not want someone to heavily assist them in completing their task.

4.3 Human-Subjects Experiment

The goal of this experiment was to gather feedback on what features needed to be incorporated into the final autonomous system. Christopher Wickens discussed some common research methods that prove effective in human factors engineering [Wickens]. He stated that when conducting an experiment, a researcher must define the problem and hypothe-

sis, specify the experimental plan, conduct the study, analyzing the data, and finally draw conclusions about the results.

Wickens also pointed out that another important aspect of an experiment is the design, which can consist of either a two-group, factorial, between-subject, and other designs. This experiment will use a between-groups, mixed factorial design and consisted of having the participant tape a cardboard box together using only either their dominant or non-dominant hand. There were 12 participants in the experiment, all of whom were engineering students at the University of Kansas, ranging from ages 18-25. Of the 12 participants, 6 were male and 6 were female. At the researcher's discretion, 6 participants were chosen to use their dominant hand while 6 used their non-dominant hand. Additionally, 6 participants received scaffolding from the Pepper robot, while 6 did not. These details are summarized in 4.2.

Wickens also listed descriptive methods that can lead to the collection of useful experimental data, including observation and incident and accident analysis, before giving analysis techniques for descriptive measures, such as observing differences between groups, relationships between continuous variables, and complex modeling and simulation. In this experiment, participants were timed to see how long it would take them to complete the task. Upon completion of the task, participants had to fill out a survey consisting of 4 questions. More details can be found in [Moore and Williams2020].

4.3.1 Wizard-of-Oz Technique

As stated earlier, the experiment incorporates the Wizard-of-Oz (Woz) technique, as Pepper may only provide encouragement at the operator's discretion. Laurel Riek proposed new reporting guidelines to adhere to when incorporating Wizard of Oz (WoZ) techniques in HRI [Riek2012]. Riek's new reporting guidelines for WoZ are divided into four parts: robot, user, wizard, and general, with each part offering questions to be considered when conducting a WoZ experiment.

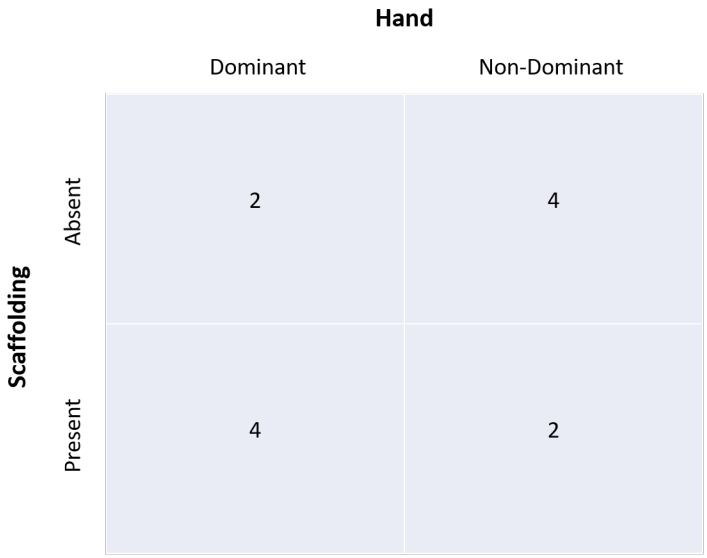

Figure 4.2: Mixed Factorial Design Setup

For this WoZ experiment, subjects could ask the Pepper robot for help during the task, at which point the robot may or may not provide words of encouragement, as the robot's responses were controlled by the researcher. The encouraging phrases included "Don't give up! You've almost got it!", "So close! You're almost there!" and "Come on! You got this!".

4.3.2 Evaluation

There has been extensive research conducted on how to evaluate HRI experiments. James Young et. al introduced a new framework for evaluating the interaction experience between humans and robots [Young et al.2011]. This framework consists of three major perspectives. Perspective 1 (P1), which deals with the visceral factors of interaction, which includes instinctual reactions of joy, fear, frustration, and other emotions. Perspective 2 (P2) observes the social mechanics of HRI, which includes gestures, spoken language, and cultural norms. Finally, perspective 3 (P3) looks at the social structure of HRI, primarily how the social relationship between a human and a robot progresses over time. The experiments conducted in this thesis will mainly focus on P1 and P2.

For this experiment, the following four hypotheses were made:

1) Subjects using their dominant hand will not find Pepper's encouragement helpful. Subjects using their dominant hand will be comfortable completing the task and will do so in a timely manner. Therefore, they will not need assistance from Pepper.

2) Subjects using their dominant hand will not vary significantly in the time needed to complete the task with or without Pepper's encouragement. Subjects using their dominant hand will feel comfortable completing the task and will do so in a timely manner. Therefore, Pepper's encouragement should not make a significant difference.

3) Subjects using their non-dominant hand will have a better experience when Pepper is providing encouragement compared to when Pepper does not provide encouragement. Subjects using their non-dominant hand will more than likely become more frustrated than subjects using their dominant hand. Therefore, Pepper's encouragement will enhance their experience.

4) Subjects using their non-dominant hand will complete the task in less time when Pepper is providing encouragement compared to when Pepper does not provide encouragement. Subjects using their non-dominant hand might feel inspired to complete the task faster if Pepper provides them with encouragement.

4.3.3 Survey Questions

The survey questions were designed to gain valuable information from participants in the experiment. The first question asked the participant to rate their experience performing the experiment on a scale from 1 to 5, with 5 being the most pleasurable experience. The second question asked the participant to give the time that it took them to complete the task. The third question asked how useful the participant felt that Pepper was during the experiment. The final question asked the user to provide additional feedback for making the AIDA system more helpful to the human user.

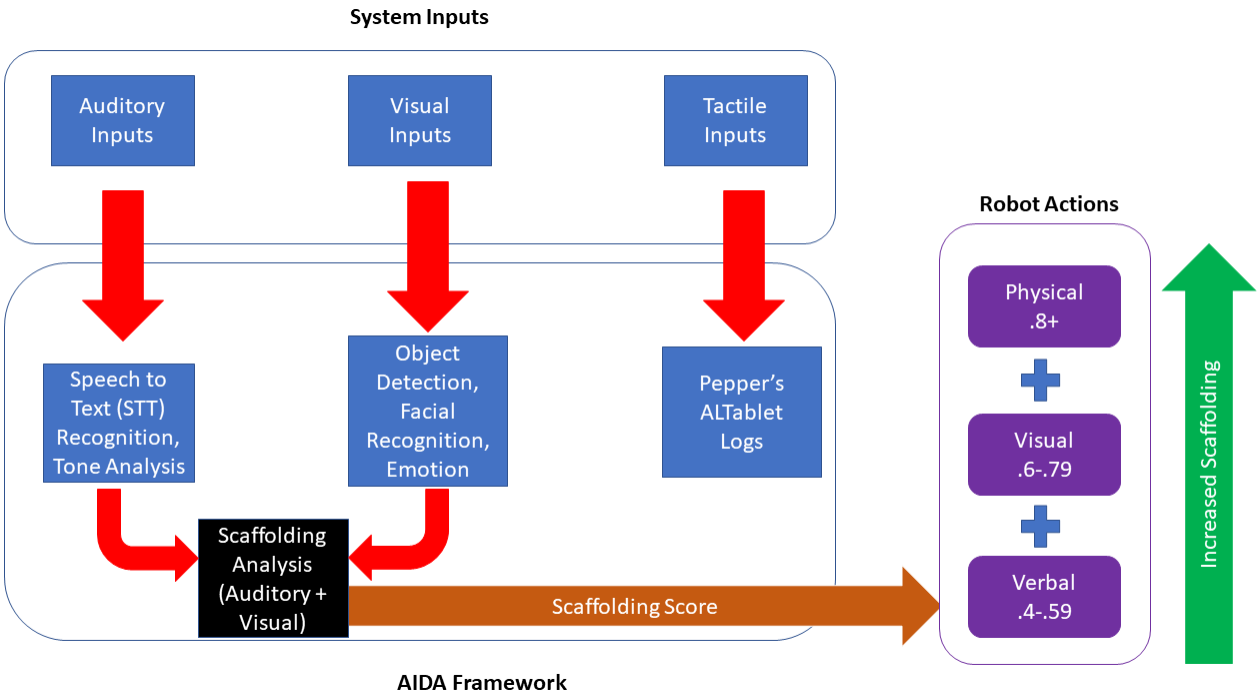

Figure 4.3: Top-level Diagram of the AIDA Architecture

4.4 AIDA Architecture

As can be seen from 4.3, the AIDA architecture consists of gathering data through the robot's inputs, processing this data through an algorithm, and finally using the output score of the algorithm to determine the robot's output actions and behavior.

4.4.1 System Inputs

As stated earlier, the Pepper robot has a variety of ways in which it can collect data from humans. The robot can gather auditory data through its four microphones, which are all located on its head. Regarding visual data, the robot can capture images and videos through its two cameras, which are located above the eyes and in the mouth. In regards to tactile inputs, the robot has a tablet that is mounted on its chest that humans and use in order to communicate to the robot in a more direct manner.

4.4.2 Algorithm for Social Scaffolding

The resulting equation will determine the output social scaffolding score for the system:

$$.5 * (emotional\ score) + .5 * (tonal\ score) = scaffolding\ score \tag{4.1}$$

For both the emotional and tonal scores, the output values will range from 0 to 1 and will reflect the five emotions. A high score indicates the need for extensive assistance from the robot. Joy will receive an output score of .2, as a person that is joyful while performing a task usually does not require assistance. Sadness and fear will be given scores of .4 and .6 respectively, as people exhibiting these emotions during a task indicate that they require moderate assistance. Disgust and anger will receive scores of .8 and 1 respectively, as these emotions are signs of great frustration with the task. If for some reason no emotion can be detected, then the output score will be 0.

The emotional score will be determined by the output of the emotion recognition model, while the tonal score will be determined by the output of the tonal analysis model. Shizhe Chen and his team constructed such a model [Chen et al.2016]. They created an emotion recognition system for video clips from three separate emotion recognition models. The first model used audio data, the second model used video data, and the last model used image data. The output scores of these three models were then fused together to determine the emotion recognized in the video clip. They achieved increased accuracy on emotion recognition when they used all three of these models versus just using one of these models.

4.4.3 Robot Outputs

Depending on the output score of the algorithm, the robot may take several different actions to assist the human in their task. If the social scaffolding score is less than .4, then the robot will not take any action to help the human, as it has been determined by the algorithm that

the human is not having any difficulties completing the task.

Verbal

If the social scaffolding score is between .4 and .59, then the robot will verbally assist the user which will be similar to the direction technique in Mascolo's research [Mascolo2005]. This involves the robot talking the human through the necessary to complete the task successfully.

Visual

If the social scaffolding score is between .6 and .79, then the robot will provide visual aids to the user in addition to verbal assistance. This is similar to the concurrent modeling and imitation technique mentioned in Mascolo's work [Mascolo2005]. This involves the robot displaying demonstrations of the task's steps on its tablet in addition to talking the human through these steps.

Physical

If the social scaffolding score is .8 and greater, then the robot will provide physical support in addition to the visual aids and verbal assistance. This is similar to the guided modeling and imitation technique mentioned in Mascolo's paper [Mascolo2005]. This involves the robot helping the human complete its tasks in addition to providing visual aids on its tablet and talking the human through these steps.

4.5 Simulated Experiment

Data collected from the initial experiment will be used to train and analyze the emotion recognition and tone analysis modules during the simulated experiment. Ashish Kapoor et. al used a similar approach to build a system to predict frustration in humans using nonverbal

Figure 4.4: Sample Images from CK+ Dataset [Lucey et al.2010]

cues as input [Kapoor et al.2007]. They argued that nonverbal cues are more useful than verbal cues since people may try to be deceitful with their word choice. To gather more insight into this hypothesis, they first conducted an experiment in which participants must try to complete the Tower of Hanoi problem with the guidance of an autonomous agent. If at any point the individual becomes frustrated with the problem, they are instructed to click either the 'I'm frustrated' or 'I need some help' buttons. After the experiments, they analyzed the nonverbal behavioral data up to the point at which the participants click the frustration or help buttons. They then used this data to train different classification algorithms including random chance, 1-nearest neighbor, support vector machine (SVM), Gaussian process, and an SVM and Gaussian process hybrid, with the Gaussian and SVM-Gaussian hybrid models achieving the highest accuracy for predicting frustration in humans.

4.5.1 Dataset Preparation

After the data from the human-subjects experiment was curated, it was realized that there were no quality images available for training the emotion recognition model. This was largely due to the orientation of the video camera during the experiment. In order to provide the model with sufficient data, the extended Cohn-Kanade (CK+) dataset created by Jeffrey Cohn, Takeo Kanade, and others was incorporated into the original dataset [Lucey et al.2010]. The CK+ dataset shows people displaying the emotions of anger, contempt, disgust, fear, happiness, sadness, and surprise. For the purposes of this research, the faces displaying contempt and surprise will be excluded from this work. Some sample images

of the dataset can be seen in 4.4.

Regarding the tonal analysis module, audio from video recordings of the experiment was extracted and manually transcribed by the researcher. The audio from the remaining clips was discarded since some participants did not attempt to speak to the robot during the experiment or their utterances were not long enough to provide quality data. These transcriptions were then provided to the module. After reviewing the videos of the 12 participants, there were 10 audio transcriptions used for testing. These transcriptions were subjectively labeled by the researcher with one of the five emotions.

4.5.2 Emotion Recognition Module Architecture

The ResNet architecture was used for emotion detection. This architecture was developed by Microsoft researcher Kaiming He and his team and submitted to the ImageNet Large Scale Visual Recognition Competition (ILSVRC) 2015 classification task, where it received first place [He et al.2015]. The ResNet framework drew its inspiration from the Visual Geometry Group (VGG) network. The VGG network is composed of a stack of convolutional layers followed by three fully-connected (FC) layers [Simonyan and Zisserman2014]. Similarly, the ResNet framework is a stack of convolutional networks followed by one FC layer. Additionally, this network has shortcut connections inserted between the convolutional layers to increase the number of parameters in the network. The 34-layer, 50-layer, 101-layer, and 152-layer versions of ResNet were tested and compared to the performance of previous state-of-the-art models such as GoogleNet and VGG, and the results showed that ResNet outperformed on the ImageNet dataset. The 34-layer version of the ResNet model can be seen in 4.5.

The 18-layer version of ResNet was selected for the emotion recognition model due to the small amount of data available for training and testing. Additionally, a pre-trained ResNet model was used for the emotion recognition model. Wan Ding and his group used

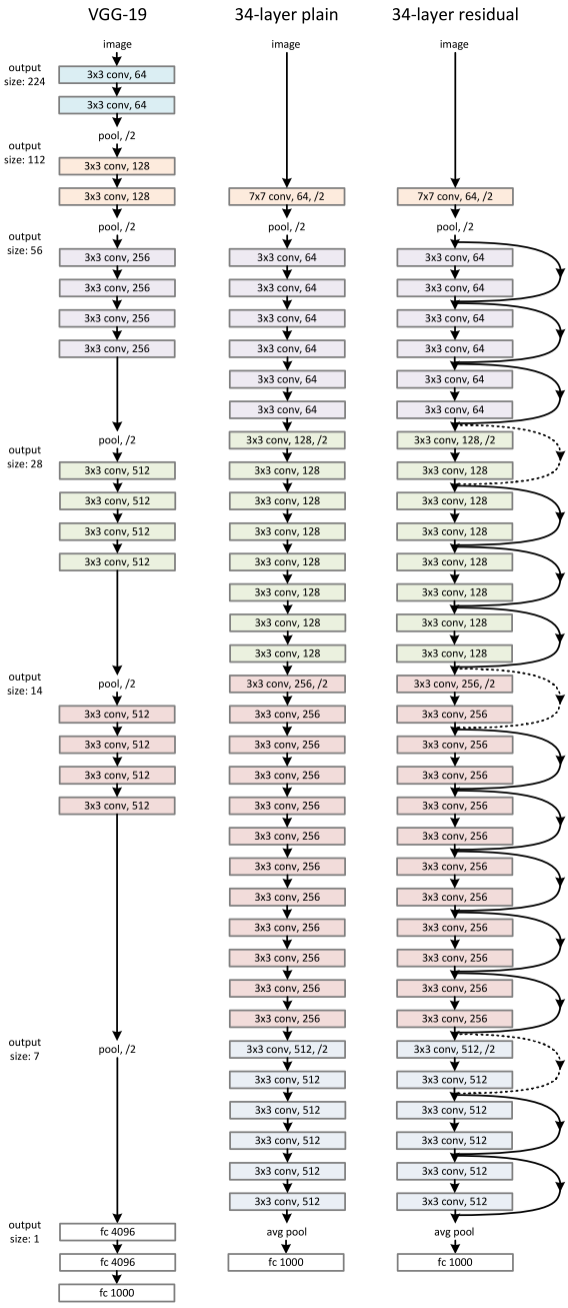

Figure 4.5: ResNet architecture from [He et al.2015]

this technique when creating their emotion recognition model [Ding et al.2016]. They used the pre-trained AlexNet model and fine-tuned the hyperparameters using the 2013 Facial Expression Recognition (FER2013) dataset.

4.5.3 Evaluation

Both the emotion recognition and tonal analysis modules were tested separately. Michael de Jong and his team separately tested their modules for their social robot system [de Jong et al.2018]. Similarly, Edmanuel Cruz and his group individually tested the modules of their assistive robot system [Cruz et al.2018].

The emotion recognition module was trained on images from the CK+ dataset for a total of 15 epochs. After tweaking the number of epochs used for training, it was found that 15 was the amount that provided the most optimal results. Due to the small number of training examples in the CK+ dataset, the model's learning converged fairly quickly. When training for longer than 15 epochs, the model began to overfit. For the tonal analysis module, the 10 prepared audio transcriptions were fed into the Watson Tone Analyzer. For a given transcription, the emotion with the highest output score was compared to the emotion label given by the researcher.

Chapter 5

Results

5.1 Human-Subjects Experiment Results

5.1.1 T-test Results

For the human-subjects experiment, none of the t-test results for the hypotheses proved to be statistically significant. For hypothesis 1, the t-test value was $p = .633$. For hypothesis 2, the t-test value was $p = .720$. For hypothesis 3,the t-test value was $p = .116$. For hypothesis 4, the t-test value was $p = .597$. Of all the t-test results, hypothesis 3 was the closest to being statistically significant.

	Dominant	Non-Dominant
	77.32	109.73
	74.71	65.97
	61.22	105.65
	73.20	78.58
	78.95	68.28
	69.01	235.66
Average Time	72.40	110.65

Table 5.1: Participants' Times to Complete Task in Seconds

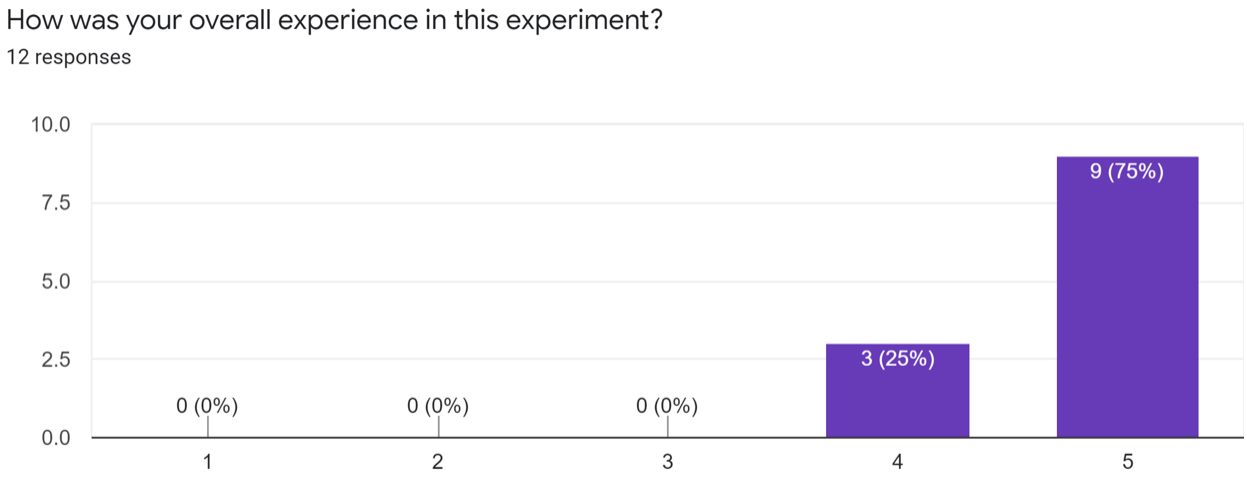

Figure 5.1: Question 1 Results

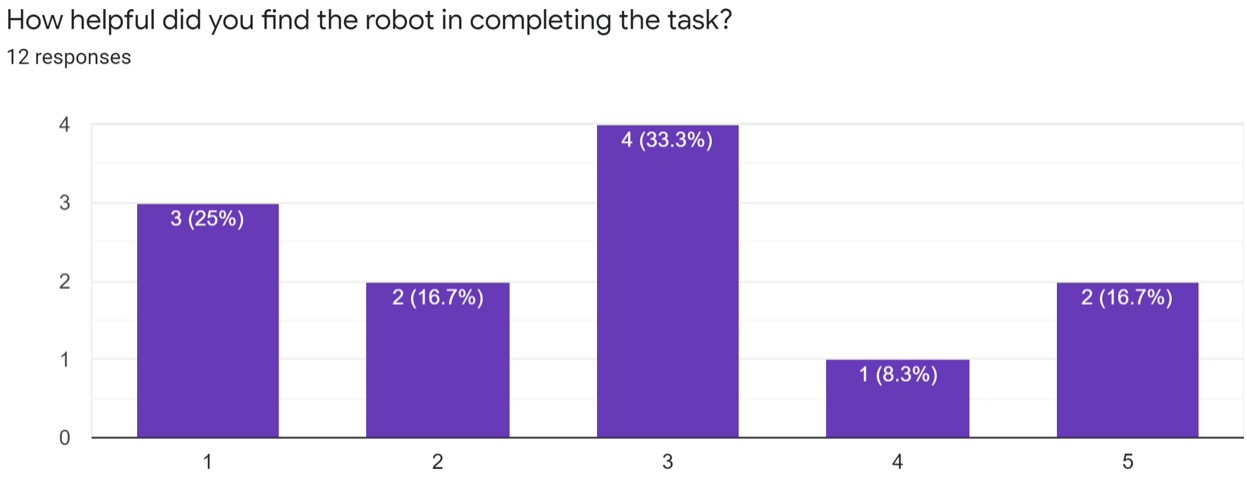

Figure 5.2: Question 3 Results

5.2 Survey Results

5.2.1 Dominant vs. Non-Dominant

In the case of dominant vs. non-dominant hand, as seen in 5.1, on average participants using their dominant hand completed the task in less time ($average\ time = 72.40\ seconds$) than

participants who were using their non-dominant hand (*average time* = 110.65 *seconds*). Additionally, participants who used their dominant hand rated Pepper as being very helpful (*average helpfulness* = 3.33), while those that used their non-dominant hand rated Pepper as not being helpful (*average helpfulness* = 1.83). In regards to the overall experience, there was little variation between participants that used their dominant hand (*average experience* = 5.00) versus those that used their non-dominant hand (*average experience* = 4.50). In general, as seen in 5.1, participants felt that they had a great overall experience. Relating to emotions, there was little difference in the number of positive emotions exhibited between participants using their dominant hand (*average positive emotions* = 2.00) those using their non-dominant hand (*average positive emotions* = 2.17). However, there was a significant difference in the number of negative emotions displayed between dominant hand participants (*average negative emotions* = 0.17) and non-dominant hand participants (*average negative emotions* = 1.83).

5.2.2 Scaffolding vs. No Scaffolding

In the case of scaffolding vs. no scaffolding, on average, participants that received scaffolding from the robot (*average time* = 77.03 *seconds*) varied significantly in time from those who did not receive scaffolding (*average time* = 106.02 *seconds*). Relating to helpfulness, participants who received scaffolding from the robot found it to be helpful (*average helpfulness* = 3.33), while those that did not receive scaffolding rated the robot as not being helpful (*average helpfulness* = 1.83). In general, as seen in 5.2, participants did not find the robot to be very helpful. Regarding overall experience, there was no significant difference in participants who received scaffolding (*average experience* = 5.00) compared to those that did not (*average experience* = 4.50). Regarding emotions, there was a considerable difference in the number of positive emotions exhibited between participants who received scaffolding from the robot (*average positive emotions* = 2.33) and those that did not

(*average positive emotions* = 1.83). However, there was not much difference in the number of positive emotions exhibited between participants who received scaffolding from the robot (*average negative emotions* = 0.83) and those that did not (*average negative emotions* = 1.17).

5.3 Simulated Experiments Results

Upon completion of training, the emotion recognition model finished with training accuracy of roughly 59.44% and a final validation accuracy of 54.35%. The tonal analysis model only predicted one label correctly from the 10 audio transcriptions. Some transcriptions were not provided output scores for any of the available emotions.

Chapter 6

Discussion

6.1 Human-Subjects Experiment Results

6.1.1 Hypothesis Results

Hypothesis 1 posited that subjects using their dominant hand will be comfortable completing the task in a timely manner and therefore will not find any assistance from Pepper particularly useful. The t-test results of this hypothesis proved to be statistically insignificant ($p = .633$). Out of the 6 subjects using their dominant hand in the experiment, 4 received encouragement from the Pepper robot. One possible explanation is that they must have enjoyed the small social interaction they had with the robot and as a result, they rated the robot as being more useful. From the 2 subjects that used their dominant hand and did not receive encouragement from the Pepper robot, one of the subjects rated the Pepper robot as not being helpful ($helpfulness = 2.00$). However, the other subject rated the robot as being helpful ($helpfulness = 2.00$) despite receiving no encouragement. A reason for this could be that this subject enjoyed the presence of the Pepper robot, even though there was no verbal interaction.

Hypothesis 2 posited that subjects using their dominant hand will feel comfortable com-

pleting the task and will do so in a timely manner and therefore the presence or absence of Pepper's encouragement should not make a significant difference. The t-test results ($p = .720$) of this hypothesis proved to be statistically insignificant. One potential reason for this is that although these subjects were using their dominant hand, their level of dexterity in the dominant hand may have varied significantly.

Hypothesis 3 posited that subjects using their non-dominant hand will more than likely become more frustrated than subjects using their dominant hand. As a result, Pepper's encouragement will enhance their overall experience compared to those who do not receive encouragement. The t-test results of this hypothesis proved to be statistically insignificant ($p = .116$). However, of all the hypotheses, this one was the closest to being statistically significant. This is to be expected, as the subjects using their dominant hand would be more comfortable completing the task and thus be less frustrated than subjects that completed the task in with their non-dominant hand. Additionally, these results show that perhaps encouragement is not sufficient enough to assist those struggling with completing a task. Scaffolding techniques that provide increased assistance to the struggling individual, such as direction and asymmetrical assistance, would need to be incorporated into the final autonomous system.

Hypothesis 4 posited that subjects using their non-dominant hand might feel inspired to complete the task faster if Pepper provides them with encouragement. The t-test results of this hypothesis proved to be statistically insignificant ($p = .597$). One possible reason for this is that no matter the amount of encouragement provided by the robot, the lack of dexterity in the non-dominant hand was too much for the subjects to overcome. However, there is some potential evidence that the absence of social scaffolding from the robot only serves to exacerbate the issues the subject faces when using the non-dominant hand. Of the 4 subjects that used their non-dominant hand and did not receive encouragement from the Pepper robot, 1 subject completed the task in a little under 4 minutes ($time = 235.66\ seconds$),

which was easily the slowest time out of all the participants.

6.2 Simulated Experiment Results

The emotion recognition module reached a final training accuracy of 59.44%. One potential reason for this is that there simply was not enough training examples for the CNN model to work with. A dataset to consider in the future is the 2013 Facial Emotion Recognition (FER2013) dataset, which contains over 35,000 images of people displaying emotions of anger, disgust, fear, joy, sadness, surprise, and neutrality. Another possible reason could be that the hyperparameters for the network were not properly tuned. However, it is not guaranteed that the performance of the model would have significantly improved.

The results show that the outputs of the tonal analysis module do not seem to be consistent with the labels created by the researcher. This seems to be due to a difference between the textual information available to the tonal analysis module and the additional audio information available to the researcher. In the future, it will be necessary to also detect emotion from audio data.

Chapter 7

Conclusion and Future Work

7.1 Future Work

7.1.1 Real-Time Testing

For future work, it is a priority to test the AIDA system in a real-time experiment. Ideally, the experiment would be designed such that Pepper had a clear view of the subject's face and could hear the subject clearly. If the results are deemed satisfactory by the researchers, then the next step will be to test the AIDA system at Cottonwood or other facilities that employ workers with IDD. Potential difficulties with on-site testing include the high variations in mannerisms among workers with different levels of IDD.

7.1.2 Increased Sociability

In future iterations of the AIDA system, it is also hoped to incorporate more social elements into the system. A more sociable robot has the ability to have more engaging and meaningful interactions with humans.

One way to improve the sociability of AIDA is to increase the robot's verbal conversation

abilities. There has been extensive research on creating more natural conversations between robots and humans. Starkey Duncan attempted to model a turn-taking mechanism for conversations between humans [Duncan1972]. He gathered data collected from the videotape recordings of two face-to-face conversations between two people. Through analysis of these conversations, Duncan made some key observations about the mechanics of turn-taking in conversations. One important feature is a turn-yielding signal that the speaker gives to the listener to let them know that they are finished speaking and the listener may begin speaking. Another feature is the speaker's use of attempt-suppressing signals to prevent the listener from attempting to take a turn speaking. Duncan observed that when the speaker gave this type of signal, the auditor's chances of attempting to take a turn speaking significantly decreased. These findings led Duncan to the conclusion that when neither the turn-yielding signals nor attempt-suppressing signals are used by the speaker, the turn-taking mechanism breaks down and leads to strange dynamics such as both parties trying to speak at the same time or both parties being silent at the same time. This research will be useful for future iterations of AIDA since turn-taking is an essential part of conversations.

Similarly, Adrian Bangerter and Herbert Clark observed how humans use dialogue to coordinate joint activities [Bangerter and Clark2003]. They believed that participants of joint activities use specific dialogue as project markers to signal horizontal transitions and vertical transitions within a joint activity. Horizontal transitions refer to the continuation of a task within a joint activity, while vertical transitions signal the entering and exiting of a task or the joint activity as a whole. More specifically, they propose that the word okay is used for vertical transitions in a joint activity, while mhm, uh-huh, yeah, and other variations of this word are used during horizontal transitions in joint activities. Analysis of transcripts from the United States confirmed their hypothesis, while analysis of transcripts from Switzerland and England did not due to a difference in language. These findings prove useful for future work with AIDA when modifying the designated output actions for the

guided modeling and imitation scaffolding technique.

Min Lee and his team conducted research on using social cues to mitigate breakdowns in robotic services [Lee et al.2010]. They conducted an experiment in which participants were placed in front of a computer and watched a scenario in which a robot takes a person's drink order and brings the person the wrong drink. However, the authors added many variables to this experiment, including the type of robot used, whether the robot offered a service recovery, and whether the robot warned the user of the difficulty of performing a service. The results of the experiments confirmed the authors' hypotheses, mainly that the participants reacted negatively to the robot's breakdown and that both forewarning the human of robot's limitations and providing the robot with a service recovery successfully mitigated the impact of the service breakdown. These findings can be used for future research when developing ways for the Pepper robot to communicate breakdowns or failures to the user.

Cynthia Breazeal investigated the role of emotion and expressive behavior in guiding the social interaction between humans and robots [Breazeal2003]. She built a framework to allow robots to recognize emotions in humans and express the appropriate emotions as a response. Her system uses drives, affective states, active behaviors to determine the emotions that the robot Kismet will express as output. The results from Breazeal's experiments prove that her framework allows Kismet to effectively recognize the emotions of humans and express emotions to humans in a clear manner. This work can be used in future research of AIDA in order to create a more engaging social interaction between users and robots.

David Johnson and Arvin Agah explored the use of multiple modalities, context, semantics, and dialog management in HRI [Johnson and Agah2009]. Their experiment results showed that the combination of these features led to more engaging HRI than a single modality did. Incorporating these features into the AIDA system could lead to a more accurate assessment of an individual's IDD level.

Ahmed Qureshi and his team used deep reinforcement learning in order to teach a Pepper

robot the norms of human social behavior [Qureshi et al.2017]. They reported that the robot was able to successfully perform basic social interaction skills after only 2 weeks of learning. These methods could be used in future versions of AIDA in order to allow the Pepper robot and create more personal social interactions with the workers.

Similar to Qureshi, Yung Gao and his team used a deep reinforcement learning model to allow a robot to learn the proper social behavior for approaching groups of people [Gao et al.2018]. Their model was able to produce more socially appropriate behaviors than state-of-the-art models.

7.2 Conclusions

There is a substantial number of individuals in the U.S. with IDD, and many of them remain unemployed. The ones that are employed work in facilities where staff members may not be able to give the amount of attention they require. This shows the need for additional staffing in these environments, including autonomous agents. This research seeks to build a system that allowed a robot to assist workers with IDD, thereby increasing their productivity. The system is comprised of data collection and processing modules, a scaffolding algorithm module, and robot action output modules. The emotion recognition and tonal analysis modules used in the scaffolding algorithm were then tested in a simulated experiment using video recordings from the initial experiment. The results give support to the hypothesis that a robot can increase the productivity of workers with IDD. In future iterations of this system, it is hoped to increase the sociability of the robot in order to produce more engaging interactions with workers with IDD and obtain additional useful information from these individuals.

Bibliography

[APA2015] (2015). *Mental disorders and disabilities among low-income children.* Mental disorders and disabilities among low-income children. National Academies Press, Washington, DC, US.

[Abadi et al.2016] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X. (2016). Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283.

[Almalky2020] Almalky, H. A. (2020). Employment outcomes for individuals with intellectual and developmental disabilities: A literature review. *Children and Youth Services Review*, 109:104656.

[Association. and Association.2013] Association., A. P. and Association., A. P. (2013). *Diagnostic and statistical manual of mental disorders : DSM-5.* American Psychiatric Association Arlington, VA, 5th ed. edition.

[Bangerter and Clark2003] Bangerter, A. and Clark, H. H. (2003). Navigating joint projects with dialogue. *Cognitive Science*, 27:195–225.

[Brault2012] Brault, M. W. (2012). *Americans With Disabilities: 2010*, page 70–131. US Census Bureau.

[Breazeal2003] Breazeal, C. (2003). Emotion and sociable humanoid robots. *Int. J. Hum.-Comput. Stud.*, 59(1–2):119–155.

[Breazeal and Thomaz2008] Breazeal, C. and Thomaz, A. L. (2008). Learning from human teachers with socially guided exploration. In *2008 IEEE International Conference on Robotics and Automation*, pages 3539–3544.

[Chen et al.2016] Chen, S., Li, X., Jin, Q., Zhang, S., and Qin, Y. (2016). Video emotion recognition in the wild based on fusion of multimodal features. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 494–500, New York, NY, USA. Association for Computing Machinery.

[Chernova and Thomaz2014] Chernova, S. and Thomaz, A. L. (2014). *Robot Learning from Human Teachers*. Morgan Claypool Publishers.

[Cruz et al.2018] Cruz, E., Escalona, F., Bauer, Z., Cazorla, M., Rodriguez, J. A., Martínez-Martín, E., Rangel, J. C., and Gomez-Donoso, F. (2018). Geoffrey: An automated schedule system on a social robot for the intellectually challenged. In *Comp. Int. and Neurosc.*

[de Jong et al.2018] de Jong, M., Zhang, K., Roth, A., Rhodes, T., Schmucker, R., Zhou, C., Ferreira, S., Cartucho, J., and Veloso, M. (2018). Towards a robust interactive and learning social robot. In M. Dastani, G. Sukthankar, E. A. S. K., editor, *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)*, pages 883–891. International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

[Ding et al.2016] Ding, W., Xu, M., Huang, D., Lin, W., Dong, M., Yu, X., and Li, H. (2016). Audio and face video emotion recognition in the wild using deep neural networks and

small datasets. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, page 506–513, New York, NY, USA. Association for Computing Machinery.

[Duncan1972] Duncan, S. (1972). Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292.

[Fukushima1980] Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202.

[Gao et al.2018] Gao, Y., Yang, F., Frisk, M., Hernandez, D., Peters, C. E., and Castellano, G. (2018). Social behavior learning with realistic reward shaping. *CoRR*, abs/1810.06979.

[He et al.2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

[Herlant et al.2016] Herlant, L. V., Holladay, R. M., and Srinivasa, S. S. (2016). Assistive Teleoperation of Robot Arms via Automatic Time-Optimal Mode Switching. *Proceedings of the ... ACM SIGCHI. ACM Conference on Human-Robot Interaction*, 2016:35–42. Edition: 2016/04/14.

[Jain and Argall2019] Jain, S. and Argall, B. (2019). Probabilistic human intent recognition for shared autonomy in assistive robotics. *J. Hum.-Robot Interact.*, 9(1).

[Johnson and Agah2009] Johnson, D. O. and Agah, A. (2009). Human robot interaction through semantic integration of multiple modalities, dialog management, and contexts. *International Journal of Social Robotics*.

[Kaplan2001] Kaplan, F. (2001). Taming robots with clicker training a solution for teaching complex behaviors.

[Kapoor et al.2007] Kapoor, A., Burleson, W., and Picard, R. W. (2007). Automatic prediction of frustration. *Int. J. Hum.-Comput. Stud.*, 65(8):724–736.

[LeCun et al.1999] LeCun, Y., Haffner, P., Bottou, L., and Bengio, Y. (1999). Object recognition with gradient-based learning. In Mundy, J., Cipolla, R., Forsyth, D., and di Gesu, V., editors, *Shape, Contour and Grouping in Computer Vision*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 319–345. Springer Verlag. International Workshop on Shape, Contour and Grouping in Computer Vision ; Conference date: 26-05-1998 Through 29-05-1998.

[Lee et al.2010] Lee, M. K., Kiesler, S., Forlizzi, J., Srinivasa, S., and Rybski, P. (2010). Gracefully mitigating breakdowns in robotic services. In *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 203–210.

[Lucey et al.2010] Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. (2010). The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 94–101.

[Mascolo2005] Mascolo, M. F. (2005). Change processes in development: The concept of coactive scaffolding. *New Ideas in Psychology*, 23(3):185 – 196. Revisiting scaffolding: What do we think we know and what do we still need to figure out?

[McKenzie et al.2016] McKenzie, K., Milton, M., Smith, G., and Ouellette-Kuntz, H. (2016). Systematic Review of the Prevalence and Incidence of Intellectual Disabilities: Current Trends and Issues. *Current Developmental Disorders Reports*, 3(2):104–115.

[Moore and Williams2020] Moore, R. and Williams, A. B. (2020). Aida: Using social scaffolding to assist workers with intellectual and developmental disabilities. In *Companion*

of the 2020 ACM/IEEE International Conference on Human-Robot Interaction, HRI '20, page 366–368, New York, NY, USA. Association for Computing Machinery.

[Pandey and Gelin2018] Pandey, A. K. and Gelin, R. (2018). A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. *IEEE Robotics Automation Magazine*, 25(3):40–48.

[Paszke et al.2019] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., dAlché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8026–8037. Curran Associates, Inc.

[Qureshi et al.2017] Qureshi, A. H., Nakamura, Y., Yoshikawa, Y., and Ishiguro, H. (2017). Robot gains social intelligence through multimodal deep reinforcement learning. *CoRR*, abs/1702.07492.

[Riek2012] Riek, L. D. (2012). Wizard of oz studies in hri: A systematic review and new reporting guidelines. *J. Hum.-Robot Interact.*, 1(1):119–136.

[Simonyan and Zisserman2014] Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition.

[Takeuchi et al.2020] Takeuchi, K., Yamazaki, Y., and Yoshifuji, K. (2020). Avatar work: Telework for disabled people unable to go outside by using avatar robots. In *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, page 53–60, New York, NY, USA. Association for Computing Machinery.

[Thomas et al.2019] Thomas, S., Suzuki, M., Huang, Y., Kurata, G., Tüske, Z., Saon, G., Kingsbury, B., Picheny, M., Dibert, T., Kaiser-Schatzlein, A., and Samko, B. (2019). English broadcast news speech recognition by humans and machines. *CoRR*, abs/1904.13258.

[Thomaz and Breazeal2008] Thomaz, A. L. and Breazeal, C. (2008). Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artif. Intell.*, 172(6–7):716–737.

[Wang and Pal2015] Wang, Y. and Pal, A. (2015). Detecting emotions in social media: A constrained optimization approach. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 996–1002. AAAI Press.

[Wickens] Wickens, C. D. Introduction to human factors engineering, 2nd edition.

[Williams et al.2019] Williams, A. B., Williams, R. M., Moore, R. E., and McFarlane, M. (2019). Aida: A social co-robot to uplift workers with intellectual and developmental disabilities. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*, HRI '19, page 584–585. IEEE Press.

[Young et al.2011] Young, J. E., Sung, J., Voida, A., Sharlin, E., Igarashi, T., Christensen, H. I., and Grinter, R. E. (2011). Evaluating human-robot interaction. *International Journal of Social Robotics*, 3(1):53–67.

Appendix A

Appendix

A.1 Post-Experiment Survey Questions and Answers

Question: **What suggestions do you have for improving AIDA's helpfulness?**

Answers

Talking to the participant during the whole process.

Nothing. She did great.

Have it give me helpful tips.

I didn't have to use AIDA to tape up the box.

She was very encouraging.

Tell the students to use pepper while they are doing the task.

N/A

Get more interaction.

If it would talk to me during closing the box.

I love the motivation it gave me, do more motivational speaking.

Idk. I didn't need help.

Have AIDA give out tips at different time intervals.