A Learning Social Referencing Disambiguation Framework for Domestic Service Robots

by

Kaiwen Fan

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2023

© Kaiwen Fan 2023

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Kaiwen (Kevin) Fan was the sole author for Chapter 1, 2, and 5 which were written under the supervision of Prof. Kerstin Dautenhahn and Prof. Chrystopher L. Nehaniv. This thesis includes first-authored peer-reviewed material that has appeared in conference proceedings, published by the Institute of Electrical and Electronics Engineers (IEEE).

Research presented in Chapters 3 and 4:

In Chapter 3, Kevin Fan designed all systems described in [11], where Dr. Melanie Jouaiti provided technical guidance. and proposed systems are integrated in the initial framework in [12] by Kevin Fan, Melanie Jouaiti, and Ali Noormohammadi-Asl. Later, the framework is extended by Kevin Fan, Melanie Jouaiti, and Ali Noormohammadi-Asl. Each component of the framework is expanded on for this thesis by Kevin Fan. The implementations of the framework and their descriptions were solely written by Kevin Fan.

In Chapter 4, the system validation study was designed and run by Kevin Fan. Melanie Jouaiti provided technical counsel. All authors reviewed and edited the manuscripts [11, 12].

Citations:

Kevin Fan, Melanie Jouaiti, Kerstin Dautenhahn, Chrystopher L. Nehaniv. Fuzzy Object Ambiguity Determination and Human Attention Assessment for Domestic Service Robots. In *2022 9th International Conference on Soft Computing & Machine Intelligence (ISCMI) pp. 140-145.* IEEE.

DOI: https://doi.org/10.1109/ISCMI56532.2022.10068479 [11]

Kevin Fan, Melanie Jouaiti, Ali Noormohammadi-Asl, Kerstin Dautenhahn, Chrystopher L. Nehaniv. A Social Referencing Disambiguation Framework for Domestic Service Robots. In *The 40th IEEE Conference on Robotics and Automation (ICRA 2023) pp. 11979-11985.* IEEE.

DOI: https://doi.org/10.1109/ICRA48891.2023.10161168 [12]

Abstract

The successful integration of domestic service robots into home environments can bring significant services and convenience to the general population and possibly mitigate important societal issues, such as care provision for older adults. However, home environments are complex, dynamic, and object-rich. It is, therefore, very probable that service robots will encounter ambiguity while interacting with household items. Moreover, service robots need to have the capability to obtain knowledge and continuously learn from the environment to ensure high adaptability and persistent functional adequacy.

This thesis presents a learning object disambiguation framework for domestic service robots that is inspired by the cognitive mechanism of social referencing and the human visual mental imagery perceptual experience. The framework allows the service robot to resolve various ambiguities in the object selection task and learn objects through bidirectional human robot interactions. The framework's technical details are explained in depth. We first describe the base framework, which consists of five functional components: the user command interface, fuzzy ambiguity determination, fuzzy human attention assessment, social referencing disambiguation, and short-term long-term memory object learning. We then explain our extended framework with the expansion of robot mental imagination and user objection absence detection to further enhance the robot's ability to handle novel objects and improve interaction robustness.

To showcase and validate our robot framework, we developed a system validation study with human participants. The framework is implemented on our mobile robot manipulator, Fetch, for our testing scenarios. Our experiment is designed to measure the success of the framework objectively and to understand the human perception of the robot with the framework. The study design, implementation, results, and discussions are illustrated.

Finally, we discuss the current limitations of the framework and summarize many valuable lessons learned in this research project. We conclude the thesis with an exploration of many promising and exciting future directions for the proposed framework. We believe this framework forms an important conceptual foundation for future service robots to become "lifelong learners" with human guidance.

Acknowledgements

I would like to first express my deepest gratitude to Professor Kerstin Dautenhahn for her immense and continuous mentorship and support throughout my research journey.

I would also like to thank Professor Chrystopher L. Nehaniv for his knowledge and insight in many enlightening discussions.

I'd also like to thank my thesis readers, professor Moojan Ghafurian and professor Gennaro Notomista, for reading my thesis and providing thoughtful questions and feedback.

My thanks go to Dr. Melanie Jouaiti for her valuable practical support on this project and her continuous friendship. Your experience and mentorship have been pivotal in my academic career.

Thank you to Ali Noormohammadi-Asl for your friendship and valued support, particularly in the many artistic illustrations of the project that made the project's idea considerably more intuitive and accessible to a broader audience.

Thank you to Negin Azizi, Steven Lawrence and members of the Social and Intelligent Robotics Research Laboratory, for the many insightful discussions and feedback.

Lastly, I am immensely grateful for all the support and love I received from my dearest family and friends.

This research was undertaken, in part, thanks to funding from the Canada 150 Research Chairs Program.

Dedication

This is dedicated to my friends, family, and my most supportive parents, Jing Su and Yansong Fan, who provided me with unbridled support in all aspects to allow me to pursue my dream.

Table of Contents

List of Figures ix

List of Tables xi

1 Introduction 1

1.1 Motivation . . . 2

1.2 Research Objectives and Challenges . . . 3

1.3 Contributions Summary . . . 4

1.4 Thesis Outline . . . 5

2 Related Work 6

2.1 Ambiguity Determination . . . 6

2.2 Attention Assessment . . . 7

2.3 Disambiguation Strategies . . . 7

2.4 Cognitively Inspired Robot Frameworks . . . 8

2.4.1 Robot social referencing . . . 9

2.4.2 Robot mental visual imagery/imagination . . . 10

3 A Learning Social Referencing Disambiguation Robot Framework 11

3.1 Framework Architecture . . . 12

3.1.1 System Overview . . . 13

3.1.2 User Command Interface . . . 14

3.1.3 Fuzzy Ambiguity Determination . . . 16
3.1.4 Fuzzy Human Attention Assessment . . . 22
3.1.5 Disambiguation . . . 26
3.1.6 Learning and Object Selection . . . 27
3.2 Extended Framework . . . 29
3.2.1 Robot Mental Imagery . . . 31
3.2.2 Second Loop of Social Referencing . . . 34
3.3 Interaction Example . . . 35
3.4 Preliminary Testing . . . 38
3.5 Discussion . . . 39

4 Fetch Robot Social Referencing System Validation 41
4.1 Interaction Scenarios and Procedure . . . 41
4.2 Participants . . . 43
4.3 Questionnaire . . . 43
4.4 Experimental Results . . . 48

5 Conclusion 52
5.1 Summary of Contributions . . . 52
5.2 Lessons Learned . . . 55
5.3 Limitations . . . 56
5.4 Future Directions . . . 57

References 59

APPENDICES 65

A System Validation Ethics Approval 66

List of Figures

3.1 The Fetch robot with a mounted camera. 13

3.2 A domestic service robot encounters ambiguity in an object selection task. The robot can only recognize generic containers, and it does not know which one is the "salt"; disambiguation and learning are required. 14

3.3 The high-level flow diagram of the base social referencing disambiguation framework. 15

3.4 User command Interface GUI, left shows the idle interface, right shows the verification interface, where the text in the text box is automatically populated by the robot, and it is editable in case if the speech detection is incorrect. 16

3.5 Ambiguity determination data flow and overview. 17

3.6 Ambiguity determination membership functions; top panels show input fuzzy membership functions (Antecedents) and the bottom panel shows the output fuzzy membership function (Consequent). 19

3.7 The setup for preliminary evaluation of fuzzy functional components. 21

3.8 Attention Assessment data flow and overview. 23

3.9 Facial key points for head orientation. 23

3.10 Eye key points for gaze assessment. 24

3.11 Eye gaze membership functions, colors in respective membership functions; top panels show input fuzzy membership functions (Antecedents) and the bottom panel shows the output fuzzy membership function (Consequent). 24

3.12 Attention estimation with heavy facial obstruction and different head postures. (a) Semi-attentive (head tilted to the left, eye gaze at the robot); (b) Fully-attentive(head and eye gaze directed at the robot); (c) Semi-attentive (head tilted to the right, eye gaze at the robot) . . 26

3.13 Robot disambiguation with speech and gestures. 27

3.14 Network overview of the social referencing disambiguation framework. Learning network with robot imagination (left), feedback network with speech analysis and head gesture analysis (right). 29

3.15 The robot does not know which container contains "salt" due to its limited knowledge base; the robot invokes imagination and learning and performs disambiguation with human feedback if the target is still unclear after imagination; Robot's imagination can provide mental guidance to accelerate the disambiguation and learning process. . . . 30

3.16 The high-level flow diagram of the extended social referencing disambiguation framework with robot mental imagery and a second loop of social referencing. 31

3.17 Extracted most salient object in the imagined scene for later analysis. 32

3.18 Examples of robot's visual mental imagery. 33

3.19 Network overview of the extended social referencing disambiguation framework. 34

3.20 Robot mental imagery of a 'green cup'. 35

3.21 Ambiguity level result, ambiguity level can be classified as "vHigh" which means very high, "high" which is high, "med" as medium, "low" as low and "vLow" as very low. 36

3.22 Gaze level after attention attraction. 37

3.23 The siamese similarity comparison: the left vertical pair has a similarity score of 0.72, and the right vertical pair has a similarity score of 0.95. The top query images are in the robot's view, and the bottom anchor images are in the robot's long-term memory. 37

4.1 Experiment setup and the Fetch Robot. 43

List of Tables

3.1 Task Ambiguity Example Outputs. 21

4.1 Adapted Technology-Specific Expectations Scale (TSES). 45
4.2 General Technology experience questions 45
4.3 Godspeed questionnaire on perceived intelligence with the 5-point semantic differential scale. 45
4.4 Robotic experience questions. 46
4.5 Adapted Robotic Social Attributes Scale (ROSAS) on robot competence. 46
4.6 Adapted Perceived Safety with Emotional State (PSE). 46
4.7 Adapted User experience questionaire (UEQ) 47
4.8 Robot predictability, human confidence, and robot dependability . . 47
4.9 5-point Likert scale. 47
4.10 10-point scale of measurement degree. 47
4.11 In examining participants' responses before and after their interaction with the system, we have reported the mean scores. Significant differences with a p-value less than 0.05 have been highlighted in bold, while items that retained their significance even after applying the Benjamini and Hochberg False Discovery Rate correction have been denoted with an asterisk (*). 49
4.12 The user experience following the interaction (5-point scale). 50

Chapter 1

Introduction

Service robots have seen significant success in many professional sectors, such as healthcare, agriculture, security, and more. In 2021, professional service robots saw a considerable growth of 37% compared to the previous year, while personal service robots, such as domestic service robots, saw a growth of 9% in contrast [32]. In addition, simple vacuuming robots or other indoor cleaning robots dominate the domestic service robot market, whereas domestic service robots that can manipulate objects and perform more complex tasks in households only occupy a fraction of the market in comparison. This is understandable, as home environments are incredibly complex, dynamic, and object-rich, even the fundamental task of robot object selection can pose extreme challenges due to the unconstrained nature of the environment [15]. Additionally, the default knowledge of the robot will be limited by the constraints imposed by the designer, resource availability, and various other factors. Furthermore, domestic service robots, unlike industrial or professional service robots, need to have superior capabilities to analyze and perform actions based on the human social environment. Where industrial or professional service robots are task-oriented and the role and reaction of humans are not focused, domestic service robots must complete the given task while considering behaviors that are understandable and appropriate for humans.

In this project, we propose a learning robot framework inspired by the cognitive mechanism of social referencing to enable a domestic service robot to perform object disambiguation and learn novel objects through natural human interactions. The framework is designed to be modular for maintainability and extensibility, and the base framework is further extended to incorporate the function of robot mental imagery (inspired by the human visual mental imagery perceptual experience) to al-

low the robot to be more adaptive to novel objects and make more natural learning decisions in general.

1.1 Motivation

With the rapid progression of technology, domestic service robots are becoming ever more autonomous, intelligent, and versatile [48]. However, for service robots operating in domestic settings where they need to interact with everyday household items, there remain many unsolved problems. In home environments, the immense quantity and variety of household objects, the unconstrained appearances of customized household items, and the incompleteness and uncertainties involved in casual human instructions can lead service robots into ambiguous states when handling objects. Service robots need to have the ability to identify the degree of ambiguity in an uncertain situation. Furthermore, the robot needs to evaluate the human partner's attention and possibly attract human attention to disambiguate effectively.

Humans are creatures of exploration and learning. The quest of pursuing knowledge begins from our first moment on earth and ends with our final breath. Humans are also creatures of collaboration; we do not have the claws and teeth of many powerful predators in nature, but we can cooperate and achieve something greater than any individual and accelerate a common goal's progression. These social and collaborative characteristics also influence how people learn; often, observing and communicating with more experienced individuals in the group can give critical insights and knowledge to the inexperienced. This method is particularly powerful when written knowledge is not available or is too slow to be used in certain situations. Social referencing is a powerful mechanism that helps humans adaptively change their behavior based on the feedback of their social interaction partners. Specifically, in human infants, social referencing is crucial to aid infants in resolving ambiguity and learning about their surroundings [4]. Social referencing has two necessary components: 1- information-seeking, where one makes requests for clarification directed at the social partner, and 2- behavior-regulatory, where one can alter one's behaviors based on the feedback received from others [7]. The successful modeling of social referencing could aid domestic service robots in functioning more effectively and naturally within the human social environment.

Despite the fact that social referencing is often employed as a means of disambiguation and discovery in humans, current computational models of social referencing in robotics often emphasize the modeling of emotional aspects [6, 45]. Current

systems only attempt to make the robot associate a simulated emotion with objects based on human feedback. Complementary to these approaches, our novel social referencing framework that focuses on the utilitarian purpose of object disambiguation could improve the usefulness and the market potential of a service robot, as the proper manipulation of common objects is critical to many practical applications of domestic service robots.

Drawing inspiration from this human perceptual experience, robots could also benefit from the generation of mental imagery. The combination of social learning and mental imagery could be highly effective for learning. Consider the description of a vibrant red apple or the image of a full moon against a starry sky. These simple cues can trigger the formation of mental images in our minds. Humans possess the remarkable capability to create mental imagery, enabling us to visualize objects, recall events, and immerse ourselves in past experiences [23]. Mental imagery extends beyond visual perception and encompasses the other major sensory systems, including proprioception and body awareness. However, humans heavily rely on their visual system to explore and comprehend the surrounding environment. This is evident in the significant allocation of cortical tissue dedicated to visual information processing compared to other senses [34]. By harnessing this ability, robots could adapt more effectively to unfamiliar situations and expedite the learning process for novel objects.

1.2 Research Objectives and Challenges

This thesis aims at proposing a learning social referencing disambiguation framework that allows a service robot to eliminate ambiguity and obtain knowledge of novel objects in object selection tasks through human interactions.

The robot framework should enable a robot to analyze the visual scene and the assigned task to determine ambiguity level, and the framework should be multimodal that allows for the analysis and production of speech and gesture signals with a human partner to form bidirectional interactions in order to resolve ambiguous states.

The research questions this project aims to answer are:

RQ1: How can we create a social referencing framework for service robots to handle object selection tasks?

RQ2: How feasible is it to develop an interactive framework for service robots that allows the robot to learn objects through interaction and experience?

RQ3: How do people perceive the robot with the proposed framework?

The framework can be divided into multiple functional components, explained in chapter 3. In addition, the extended framework is described later in chapter 3, allowing robots to harness mental imagery and perform passive social referencing. RQ1 is thus addressed in Chapter 3, while RQ2 is partially addressed in Chapter 3. The framework's system validation results, through a human-robot interaction (HRI) experiment, are illustrated in chapter 4, addressing the remainder of RQ2 and RQ3.

1.3 Contributions Summary

The contributions of this thesis can be summarized as follows:

1. Development of fuzzy decision systems that allow the robot to analyze the object selection task ambiguity based on user verbal or written commands and the robot's visual scene, and estimate the human attention based on human gaze and head orientation.
2. Development of and the integration of fuzzy inference systems with the user command interface, disambiguation module, long-term short-term memory object learning, and robot imagination functional block to establish a social referencing disambiguation framework[1] for domestic service robots. The framework allow the service robot to:
   (a) analyze the ambiguity level of a given task;
   (b) assess the human's attention and attract human attention when necessary;
   (c) disambiguate references to objects using human feedback (gestures and speech);
   (d) learn novel objects via robot mental imagination or human clarification.
3. Verification of the proposed framework with an in-person HRI study, along with the analysis of the experimental results, and critical discussions.

[1] https://github.com/KevinFan9729/socialReferencingDisambiguation

4. Supplemental discussions on possible future extensions of the framework so that the framework can be more generalized, more adaptive and more effective in accomplishing the vision of an intelligent general domestic service robot.

1.4 Thesis Outline

The thesis continues with chapter 2 detailing the essential background information on ambiguity determination, human attention assessment and existing related robotic systems.

Subsequently, we introduce chapter 3 to provide information on the architecture of the framework, and explain technical details and implementation details of the framework by explaining all its functional components and extensions.

Chapter 4 documents a system validation Human-Robot-Interaction (HRI) experiment where the framework runs on a mobile manipulator robot, Fetch. We show the design of the experiment, the analysis of the experimental results along with critical discussions.

Lastly, chapter 5 concludes this thesis with the summary of the project, along with the a discussion of contributions of this work, and future directions and possible extensions of the proposed framework to enable a new generation of intelligent universal domestic service robots in the near future.

Chapter 2

Related Work

This chapter provides relevant background information of related projects to explain the context and motivation for developing a human inspired social referencing disambiguation framework for domestic service robots. We provide a brief overview of the related methods and techniques on ambiguity determination, attention assessment and object disambiguation along with some related robotic frameworks.

2.1 Ambiguity Determination

In regard to ambiguity level in object selection tasks, current literature primarily utilizes simple Boolean logic for determination. In work presented by Hatori et al. [18], the system determines if a detected object's confidence level exceeds a hard threshold to be considered a potential target, and the system will determine that the situation is ambiguous when there are multiple potential targets. Similarly, the system introduced by Sibirtseva et al. [43] computes ambiguity by identifying the number of objects the system can detect based on the request of the human partner. The system determines that there is ambiguity if there are multiple potential target objects in the detection. Similarly, Haasch et al. [16] also determine ambiguity based on the number of objects that may satisfy the user request.

2.2 Attention Assessment

Fuzzy logic inference systems introduced by Zadeh [50, 51] address uncertainties in many real-world problems and closely resemble the human decision-making process, yet can maintain relatively low computational demands and good flexibility for rapid adaption to incorporate new information if necessary. Robot object selection tasks can include different kinds of uncertainty, and determining human attention involves intrinsic vagueness, considering natural daily motions of the human partner (not constrained by experimental setups). Therefore, fuzzy logic is well suited for handling complexities in human-robot interaction applications, increasing system robustness and integrating multiple data streams.

Assessing human attention is highly critical for many robotic applications. Lemaignan et al. [24] described a system that utilizes the keypoint-based head pose of the user to approximate gaze direction and subsequently uses head pose to estimate attention focus. The system computes the "with-me-ness" value, a normalized ratio metric of human attention based on the attention time on the target task and the expected interaction time. Minguillon et al. [33] proposed the detection of human attention in multi-speaker scenarios based on EEG data with fuzzy logic inference. Whereas the work introduced by Li et al. [25] relies on gaze direction measurements with fuzzy logic to evaluate a surgeon's attention and direct a robotic laparoscope system to change viewpoints during surgeries adaptively. Asteriadis et al. [3] utilize convolutional neural networks (CNN) to classify head orientation. They approximate the gaze direction by transforming eye images and computing the gaze vector by comparing the transformed eye image with the front-facing eye image. The head pose and gaze data are combined with a fuzzy inference system to calculate an attention level.

2.3 Disambiguation Strategies

Object selection is one of the most common and fundamental tasks that a service robot needs to achieve. The resolution of ambiguity involved in object selection has, therefore, been extensively explored, by exploiting various human signals, such as gestures or speech.

Gesture is one of the most commonly used methods for disambiguation [16, 46, 47]. Early work, such as [16] utilized the visual appearance of objects (color and shape)

for discrimination, and the system relied heavily on deictic (pointing) gestures from the human demonstrator to disambiguate similar objects.

Although human gestures can be interpreted relatively easily and have proved to be successful [16, 46], more recent systems rely more on natural language, due to the rising popularity and capability of deep learning. The system proposed in [47] formulated the object selection task as a Partially Observable Markov Decision Process (POMDP) that accepts human deictic gestures and speech to attempt to determine the object desired by the user. The system can ask for verbal clarification when the desired object cannot be determined from the initial human desire estimation. The system in [43] combined speech understanding of user instruction and visual display of the potential targets to allow the user to provide additional information.

The work introduced in [30] analyzes detailed natural language instructions with a multimodal target-source classifier model with attention branches. The model takes in visual information, spatial locations, and linguistic instruction. The likely target-source pairs (pairs of location of the target item and the item) are then generated, using an attention map and the user could select the most desirable pair for the robot to pick up without ambiguity.

Furthermore, the interactive dialogue system in [18] analyzed unstructured text or speech instructions with Long Short-Term Memory (LSTM) recurrent network and visual data with Single Shot MultiboxDetector and a Convolutional Neural Network. The system disambiguates references by providing speech queries to the user. Similarly, in [42] the system could analyze unstructured speech instructions and disambiguate by providing more informative queries to the user. In [42], the system performed a relevancy clustering operation to extract the target object after the LSTM and CNN analysis. Zhang et al. suggested an improved interactive dialogue system by using a POMDP model to track the history of observations and ask appropriate informative clarification questions to disambiguate [52]. More similar language-based interactive dialogue systems are discussed in [36], [35] with improved language understanding and more detailed disambiguation queries.

2.4 Cognitively Inspired Robot Frameworks

Our framework is heavily inspired by human cognitive mechanisms and perceptual experiences. Simulating human functions can facilitate robots' ability to produce more human-understandable and acceptable behaviors; thus, many robotic frameworks and systems also draw inspiration from human cognition. To conclude the

related work section, we provide a brief overview of some of the noteworthy human cognition inspired robot computational frameworks that are in the realm of social referencing and mental imagination.

2.4.1 Robot social referencing

Previous attempts have been made to simulate the social referencing mechanism in robotics. In the work proposed by Thomaz et al. [45] and the work introduced by Boucenna et al. [6], systems were proposed that incorporated emotion recognition and the robot's internal states to explore the relationship between emotions and objects. The work in [45] developed a framework on an anthropomorphic robotic platform, "Leonardo". This framework enabled the "Leonardo" robot to track 22 affectively salient facial nodes via environmental cameras and onboard cameras. The facial analysis allowed the robot to estimate emotions based on facial expressions, and it also enabled the robot to understand the head orientation of the user. The framework was capable of responding to vocal inputs via a simple word-spotting mechanism, so that the robot could understand the emotional values of short phrases based on emotive keywords. The human demonstrator could show various objects with different facial and verbal expressions, and based on these stimuli, the robot could pay attention to objects and associate the emotional contents of the demonstrated objects. In later interactions, the robot could display proper motor expressions towards objects that were associated with particular emotions. For example, the robot could express a 'fearful' expression toward a 'bad' object.

Similarly, in [6], the proposed framework was implemented on a custom robot head and manipulator combination. Through facial expression analysis and emotional modeling, the robot could understand the simple affective states of the user. To obtain focus points on objects, attention saliency maps were then computed as the combination of the color detector (color saliency) and Gabor filter (texture saliency) outputs. The system enabled the robot to observe the facial expressions of its human interaction partner and the objects presented by the partner. Based on the facial expressions of the interaction partner, the robot is able to associate internal emotions with the presented objects and trigger appropriate motor actions when encountering the objects again. For instance, if an object is associated with negative emotions, the robot arm could try to avoid it, while it would approach objects associated with positive emotional values.

2.4.2 Robot mental visual imagery/imagination

Utilizing visual mental imagery or imagination can greatly enhance a robot's overall adaptability. In earlier research conducted by Roy et al. [39], a mental imagery system incorporating spatial language understanding was introduced. This system enabled a robot manipulator to maintain object permanence and comprehend perspective differences in conjunction with a conversational agent. The mental model was created using simple 3D rigid body objects within a simulated environment, and a physical dynamic predictor was employed to anticipate the object's location in the subsequent image frame.

More recent advancements, as demonstrated in the work of Mania et al. [31], have implemented a 3D virtual reality engine and a realistic physics simulation to establish the robot's mental representation of the workspace. The approach facilitated a "general comprehension" of the objects that the robot was required to interact with, allowing for comparisons between imagined expectations and the actual sensory input. The physics simulation allowed the robot to reject physically infeasible object manipulation plans and update its expectations in the robot mental representation. Furthermore, Lin et al. [27] proposed a mental imagery model that constructs a 3D representation of the scene using multiple 2D images captured by the robot's sensors. The work here emphasized the prediction of unobservable viewpoints via Neural Radiance Fields as virtual cameras. The virtual cameras could then be used to allow a robot to 'imagine' unseen view points of an object if the robot would mentally rotate the object or reposition itself. These advancements in mental imagery models contribute to the robot's ability to perceive and interpret its environment effectively.

Our framework integrates social referencing and mental imagery techniques for disambiguation purposes, leveraging both verbal and nonverbal cues. Moreover, the framework facilitates the robot's acquisition of knowledge through interactive experiences. The incorporation of the mental imagination process enhances the robot's capabilities when dealing with objects where its base knowledge is insufficient. This is achieved by optimizing the questioning process (asking efficient questions), or even bypassing the need for interactive disambiguation entirely when the imagined object closely resembles the target item.

Chapter 3

A Learning Social Referencing Disambiguation Robot Framework

In this chapter, we describe the development and implementation details of the learning social referencing disambiguation framework for domestic service robots. The framework is highly modular and can be divided into five fundamental functional components:

- User command interface
- Fuzzy ambiguity determination
- Fuzzy human attention assessment
- Social referencing disambiguation
- Short-term long-term memory object learning.

These major five functional components form the basic proposed framework, and thanks to the conceptual modularity of the framework, we then extended the framework with two additional auxiliary functional components:

- Robot mental imagery/imagination
- Objection absence detection.

We explain the base framework along with all the fundamental functional components first, the we can elaborate on the extended framework.

The proposed framework combines both human speech and gesture feedback and utilizes simple robot speech queries and robot gestures to disambiguate the correct reference. In addition, the system also incorporates a short-term/long-term memory learning scheme to learn novel objects after disambiguation, so that the robot can execute the task without ambiguity, if the same instruction is given again. The framework draws inspiration from the human cognitive mechanism of social referencing, and it is later expanded with the concept of mental imagery to allow effective and natural human robot interactions.

The full framework is implemented on a Fetch mobile manipulator robot, as illustrated in Figure 3.1. The Fetch robot shows satisfactory mobility, object handling capability, and real-time perceptual sensing ability, making it suitable as a hypothetical domestic service robot. A Microsoft Surface laptop is used to provide a keyboard input channel and display speech queries detected by the robot. An additional Google Pixel 4a camera is mounted onto the head of the robot to provide an additional high-resolution video stream to the attention analysis block (This camera is needed as the robot's onboard camera is used for object detection, and the frame rate of the onboard camera is not very high. For attention analysis, the eye movement needs to be tracked in a high frame rate video stream.). The framework is integrated with the robot operating system (ROS). A remote Ubuntu 18.04 PC with an NVIDIA RTX 2060 GPU and an i7 Intel Core CPU (11th Gen Intel i7-11700 @ 2.5GHz x 16) hosted all data processing pipelines, deep learning models, and robot planning and control models.

3.1 Framework Architecture

Let us consider the scenario depicted in Figure 3.2 where the human user instructs the service robot to "pass the salt". The robot then looks for objects on the table and finds multiple containers. However, without a specifically trained object detection model, it cannot understand which one is the salt container. Additionally, the robot is encountering ambiguity due to detecting multiple potential targets. After determining the presence and the degree of ambiguity, the robot needs to seek clarification from the human partner to disambiguate. The robot can then proceed to execute the task, adapting its behaviors based on the human's feedback. Importantly, if the ambiguity stems from the novelty of an object, the robot needs to learn that object's

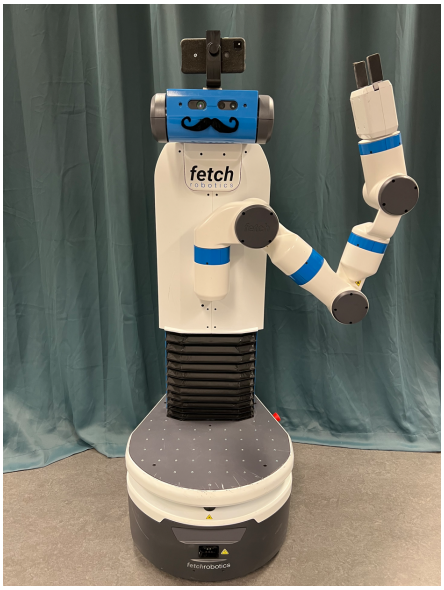


Figure 3.1: The Fetch robot with a mounted camera.

characteristics to avoid ambiguity in the future.

3.1.1 System Overview

The framework[1] (Figure 3.3) takes in natural language commands or text-based input for task instruction. It extracts the object label plus any descriptive terms of the object, e.g., "*salt* container" and "*blue* cup". The framework then analyzes the visual scene of the environment, determines if the object is known and computes the ambiguity level of the task via fuzzy logic [11]. If the ambiguity level is 'high' (i.e., there are multiple target candidates or the requested object is unknown), the framework notifies the user about the ambiguity via speech and assesses human attention level, based on visual data, and it tries to attract human attention via speech or robot gesture (i.e., waving) before the disambiguation process. The framework allows the robot to perform deictic gestures, gaze communication, and simple verbal queries to inquire about the potential target candidate. The human's speech and head gestures are analyzed to provide feedback and to enable the framework to obtain the correct reference. If the primary ambiguity source originates from the novelty of the object

[1]Source code available:
https://github.com/KevinFan9729/socialReferencingDisambiguation

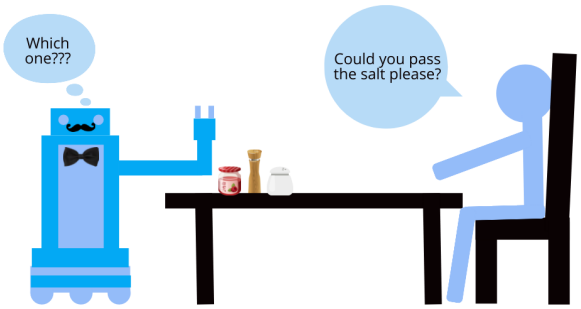


Figure 3.2: A domestic service robot encounters ambiguity in an object selection task. The robot can only recognize generic containers, and it does not know which one is the "salt"; disambiguation and learning are required.

label, the framework will store the clarified item's visual example into the long-term memory for future retrieval. Lastly, the robot fetches the determined target item.

The social referencing disambiguation framework integrates multiple data-driven, deep-learning solutions, fuzzy logic inference and expert systems to perform task, speech, and gesture analysis. The high-level flow diagram of the entire framework is shown in Figure 3.3.

3.1.2 User Command Interface

The user can provide instructions to the robot with either text-based inputs or verbal commands. To support text-based inputs and input verification, we developed an interactive graphical user interface (GUI), which is illustrated in Figure 3.4.

If the user chooses to directly issue verbal commands, we utilize the Google Speech Recognition API to transcribe speech to text. Since human language is creative and productive in nature, there are various ways to express the same idea with different vocabulary and sentence structures. For example, if the user wants the robot to pass a cup, one may say, "please pass me the cup", or "the cup please", or "hand me the cup". It is also natural for users to add descriptors to the target object such as "give me the red plastic cup". Since we do not restrict how the user issues verbal commands, the possibilities are almost limitless. Instead of building complex language models to extract the relevant information, we tokenize the entire instruction into word tokens and apply the HunPos tagger [17] to classify the tokens' word classes. We then combine all adjectives followed by nouns to construct the target label. For

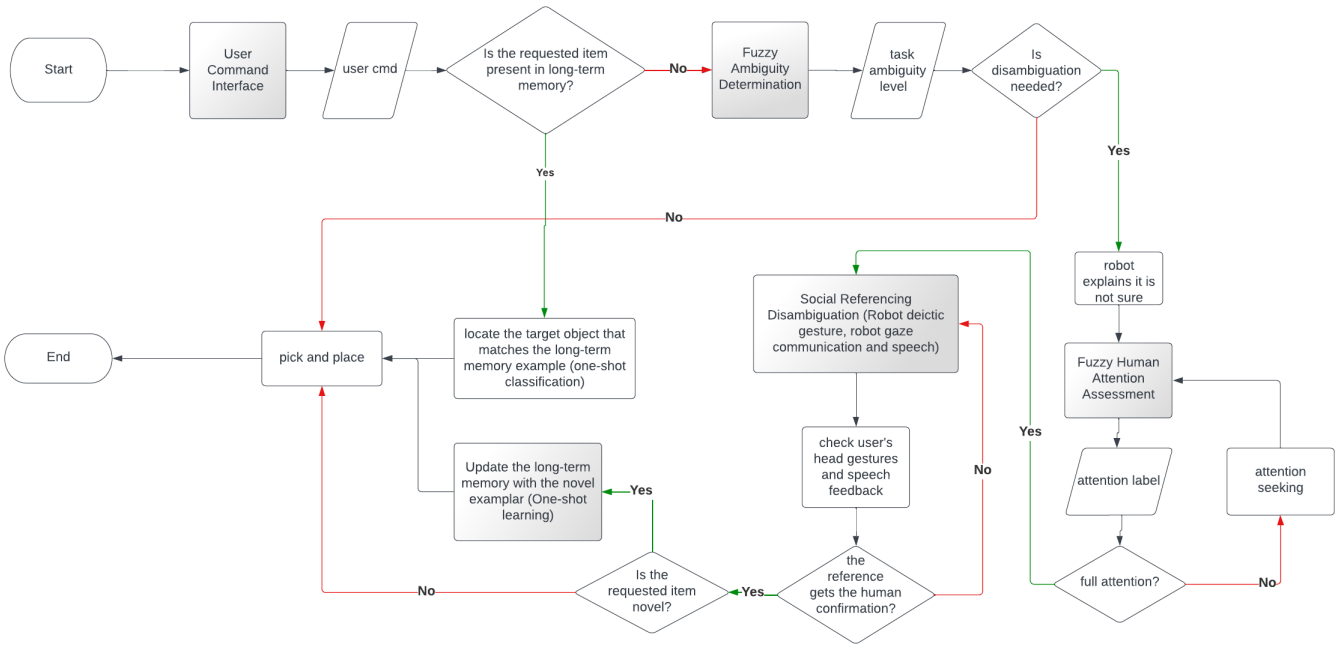


Figure 3.3: The high-level flow diagram of the base social referencing disambiguation framework.

example, for the command "can you please give me the red plastic cup" the extracted label will be "red plastic cup". Additionally, we recognize that speech-to-text transcription can be faulty due to many factors, such as accents, signal strength, and audio noise. The GUI displays what the robot heard automatically if a valid instruction is recognized, the GUI can then prompt the user, and the robot verbally asks for confirmation before processing the (potentially faulty) command. The inclusion of this additional user confirmation step intends to address the unreliability of natural language interactions with robots due to noise, weak signal strength, and user accent.

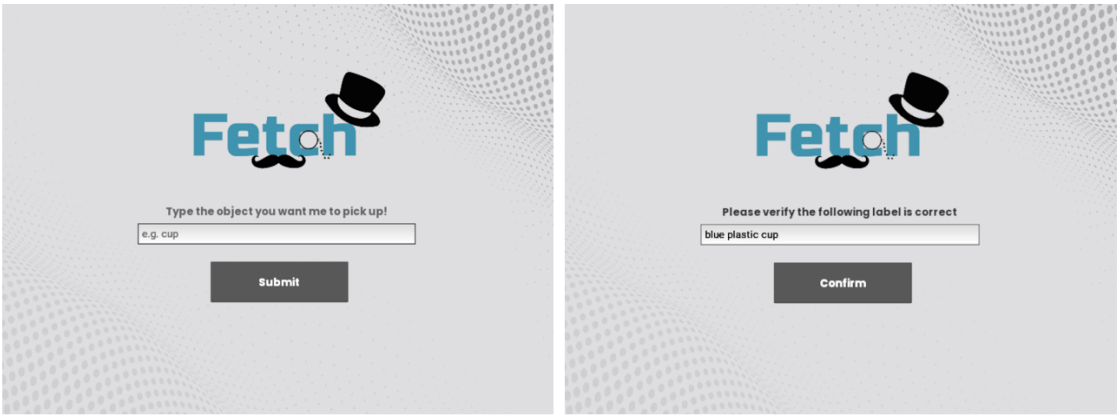


Figure 3.4: User command Interface GUI, left shows the idle interface, right shows the verification interface, where the text in the text box is automatically populated by the robot, and it is editable in case if the speech detection is incorrect.

3.1.3 Fuzzy Ambiguity Determination

After obtaining the object label, the robot needs to analyze its visual scene and determine if the current task is ambiguous. The system receives commands from the human user to instruct the robot on the item of interest. The robot utilizes the onboard RGB camera and YOLO [37] object detection trained with Microsoft COCO dataset [26] to detect objects and create bounding boxes. We then extract the object confidence of the target object(s) and the number of items that satisfy the user command. These crisp numerical values feed the fuzzy inference system [1] for analysis and data integration. Additionally, the this fuzzy system implicitly analyze the object novelty (novel objects have detection confidence of 0). Finally, the system computes a crisp numerical ambiguity level metric for the robot to determine the ambiguity of the current situation, and the robot can react accordingly based on the level of ambiguity. An overview of the fuzzy ambiguity determination block[2] is illustrated in Figure 3.5.

As described in Zadeh's fuzzy set theory and fuzzy inference system [49–51], the fuzzy logic models the imprecision of data by taking the linguistic variables as inputs and describing them with membership functions. A network of if-else rules specifies the relationships between terms in the fuzzy set, and lastly, the system produces a crisp numerical value through a defuzzification method. Since we deal with

[2]Source code available:
https://github.com/KevinFan9729/FuzzyAmbiguityandAttentionDetermination

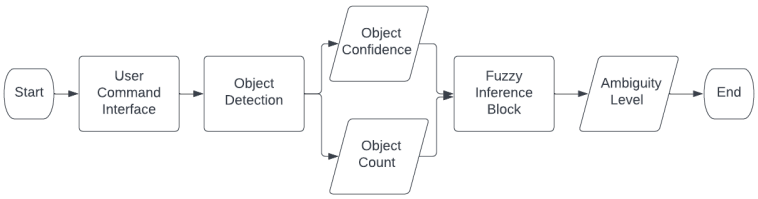


Figure 3.5: Ambiguity determination data flow and overview.

dynamic environments, language-based instructions, and real-time object detection, fuzzy logic is applied here to mimic the human decision process to handle the imprecision of real-world data. There are different kinds of ambiguity involved in the robot object selection tasks:

- Ambiguity due to multiple target candidates
- Ambiguity due to misclassification and low confidence in the object detection algorithm
- Ambiguity due to object novelty.

Instead of just giving a binary ambiguity analysis (ambiguous vs. non-ambiguous), our fuzzy logic inference system can compute a continuous ambiguity level to enable the robot to adapt to those levels and address different kinds of ambiguity.

Our fuzzy decision system utilizes the object confidence score of the target object and the number of objects that satisfy the human instruction as antecedents, and the consequent of the system is the ambiguity level of the task. The linguistic variable 'object confidence' is divided into five folds: very low (*vlow*), *low*, medium (*med*), *high*, and very high (*vhigh*); the universe of discourse of this variable is defined as a normalized measure in the numerical range from zero to one. The 'ambiguity level' variable can be considered as a percentage similar to the confidence level; thus, the ambiguity level has the same linguistic categories and the universe of discourse as the 'object confidence' variable. The 'object count' input has four terms: no object (*noObj*), single object (*one*), two potential targets (*two*), and multiple objects (*more*). The universe of discourse of 'object count' can be defined from zero to seven.

Ambiguity Determination Membership Function

The membership functions for the object confidence level and the ambiguity level are defined by Gaussian functions with means of 0, 0.25, 0.5, 0.75, 1 for *vlow*, *low*, *med*, *high*, *vhigh* respectively, and a standard deviation of 0.1, the maximum of all Gaussian functions is scaled to 1:

$$f(x) = \exp\left(-\left(\frac{x-\mu}{\sigma}\right)^2\right), \tag{3.1}$$

Ambiguity can arise due to low confidence in object detection. Low confidence can be caused by various factors such as poor lighting conditions, unusual appearances of custom objects, and more.

The object count has a membership function defined by a combinations of triangular functions, a Dirac delta function, and a trapezoidal function;

the triangular function in the general form of:

$$f(x) = \begin{cases} \frac{x-a}{b-a}, & a < x < b \\ \frac{c-x}{c-b}, & b < x < c \\ 0, & x \leq a \text{ or } x \geq c, \end{cases} \tag{3.2}$$

the translatable Kronecker delta function:

$$\delta(x-T) = \begin{cases} 1, & x = T \\ 0, & otherwise, \end{cases} \tag{3.3}$$

and the half trapezoidal function which can be specified:

$$f(x) = \begin{cases} \frac{x-a}{b-a}, & a < x < b \\ 1, & b \leq x. \end{cases} \tag{3.4}$$

together the aggregated membership functions:

$$f(x) = \begin{cases} -x+1, & 0 \leq x < 1 \\ \delta(x-1), & x = 1 \\ x-1, & 1 < x \leq 2 \\ \text{-x+3 and (x-2)}, & 2 < x \leq 3 \\ 1, & x \geq 3 \end{cases} \tag{3.5}$$

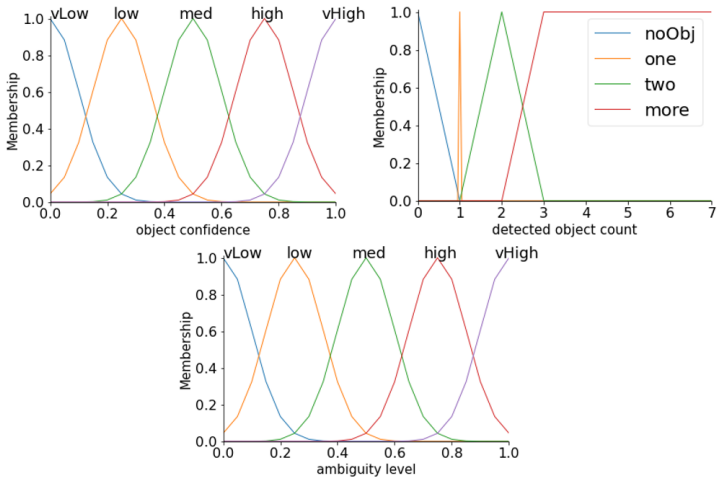


Figure 3.6: Ambiguity determination membership functions; top panels show input fuzzy membership functions (Antecedents) and the bottom panel shows the output fuzzy membership function (Consequent).

It is important to include the object count data as this variable entails a different type of ambiguity that multiple objects are potential targets. Though the fuzziness of the object count is limited, the membership function described in equation 3.5 can incorporate and integrate the object count data into the overall fuzzy inference system.

A graphical illustration of the membership functions utilized are shown in Figure 3.6. Because the fuzzy membership function describes the degree of truth that an element belongs to a subset (partial truthiness), the maximum value of the membership function is 1. The antecedent membership functions describe the visual scene, and the consequent membership function describes the overall ambiguity level.

Ambiguity Determination Knowledge Base

A network of if-else statements constructed the knowledge base of the fuzzy ambiguity unit to infer decisions. The inference rules describe the relationships between terms of linguistic variables. The ambiguity determination system utilizes eight understandable inference rules to estimate the ambiguity level of the task. The inference rules are summarized as follows:

1. **IF** (*object confidence* is very low **OR** low) **AND** (*object count* has no object **OR** multiple potential targets) **THEN** *ambiguity level* is very high

2. **IF** (*object confidence* is very low **OR** low) **AND** (*object count* is one) **THEN** *ambiguity level* is high

3. **IF** (*object confidence* is low **OR** medium) **AND** (*object count* has multiple potential targets) **THEN** *ambiguity level* is high

4. **IF** (*object confidence* is low **OR** medium) **AND** (*object count* is one) **THEN** *ambiguity level* is high **OR** medium

5. **IF** (*object confidence* is medium **OR** high) **AND** (*object count* has multiple potential targets) **THEN** *ambiguity level* is medium

6. **IF** (*object confidence* is medium **OR** high) **AND** (*object count* is one) **THEN** *ambiguity level* is low

7. **IF** (*object confidence* is high **OR** very high) **AND** (*object count* has multiple potential targets) **THEN** *ambiguity level* is medium

8. **IF** (*object confidence* is high **OR** very high) **AND** (*object count* is one) **THEN** *ambiguity level* is very low

As shown, our fuzzy rules are intuitive in describing events (a highly explainable system). Furthermore, rules are also flexible for modifications and additions when new domain knowledge is obtained. The ambiguity level is a crisp real number that is defuzzified using the centroid method described in [1], where the center of mass of the activated membership function is calculated as the output value.

Example Output

We obtained preliminary example outputs of this functional components with the Fetch robot. The user supplied the robot with the object label, and the robot utilized the proposed fuzzy system to analyze the task ambiguity.

With the Fetch robotic platform running YOLO object detection, the robot can determine the ambiguity of the object selection task with the proposed fuzzy inference. Table 3.1 summarizes the robot's determination of task ambiguity where the human partner instructs the robot to hand over an object from different objects setups on a desk.

With the evaluation of the task ambiguity on a numerical scale, the robot can decide that the current situation is clear if the ambiguity level is low. Importantly,

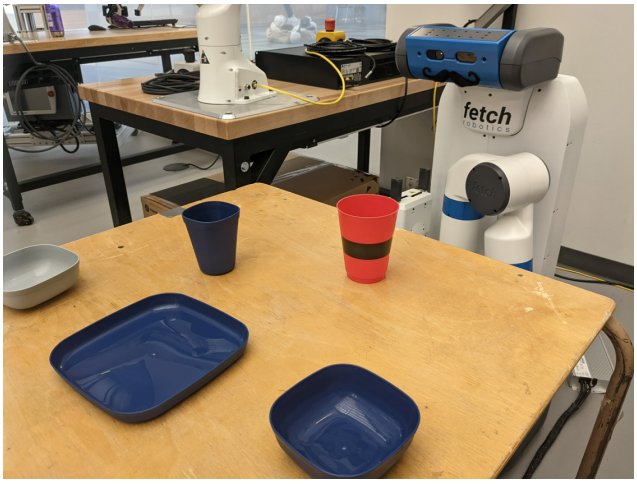


Figure 3.7: The setup for preliminary evaluation of fuzzy functional components.

the robot can be adaptive and initiate different disambiguation plans based on the various degrees of ambiguity.

Fuzzy Ambiguity Determination			
User object label	Potential Object Confidence	Num. of Objects that Satisfy the Task Instruction	Task Ambiguity Level
cup	0.93	1	0.17
bottle	0.85	1	0.21
banana	0.73	1	0.26
apple	0.64	1	0.43
cup	0.91	2	0.50
box	0.34	1	0.54
bowl	0.54	3	0.62
white cup	0	0	0.80

Table 3.1: Task Ambiguity Example Outputs.

3.1.4 Fuzzy Human Attention Assessment

The system cannot disambiguate effectively if the human user is not paying adequate attention to the robot. Therefore, the framework needs to assess the human's attention level before engaging in any disambiguation process. The system utilizes both head orientation and gaze direction to evaluate the attention level of the user. The framework applies the method proposed in [29] to extract user facial key-points, iris coordinates and performs Perspective-n-Point (PnP) solving with Levenberg-Marquardt optimization [10]. The head pose can then be determined by the thresholding logic of the head rotational values in the x and y dimensions. To estimate eye gaze direction, we apply another fuzzy inference system due to the fact that eyes are small visual features that can be noisy in terms of detection. In addition, eye movements are also considerably more subtle compared to head rotation, and the fuzzy inference system can aid the estimation to be more robust. The framework employs Gaussian membership functions in the form of equation (1) to describe the gaze direction in each eye and the activated memberships in each eye are then projected to the consequent Gaussian membership function for overall gaze direction estimation. The membership functions in each eye are also continuously calibrated with the positions of the respective eye's inner and outer corners to accommodate different eye shapes and sizes that can vary due to head movements. The eye gaze direction identification fuzzy interference system's universe of discourse is given by measurements of eye position adjusted based on local eye corners and head pose labels. The iris locations of the eyes are fed into the fuzzy system to generate an overall numerical eye gaze measurement, and the gaze labels are computed with thresholding functions adjusted with the head pose label. Finally, the human attention label can be determined with the head pose and eye gaze. Figure 3.8 provides an overview of the fuzzy human attention assessment functional component[3].

As previously described, the attention assessment system utilizes facial key points to determine head pose and eye gaze labels. Due to safety considerations of COVID-19, the system needs to detect head orientation with obstructed facial features (users may need to wear facial masks to prevent contraction of airborne COVID-19); therefore, the system utilizes key points on the forehead, the center point between the eyebrows, and two outer corners of the eyes for the head pose detection (see Figure 3.9). The head pose directional labels (head up, head down, head left, head right) can be determined with simple thresholding logic on head rotation.

[3]Source code available:
https://github.com/KevinFan9729/FuzzyAmbiguityandAttentionDetermination

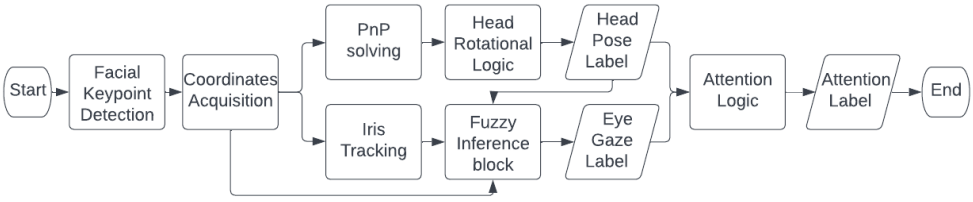


Figure 3.8: Attention Assessment data flow and overview.

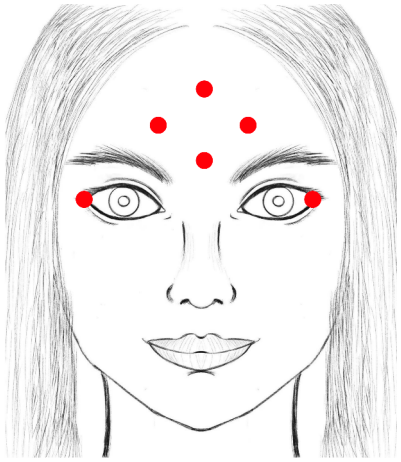


Figure 3.9: Facial key points for head orientation.

Eyes are small facial features and require special attention to process accurately and reliably. Therefore, a fuzzy inference system is implemented that incorporates head rotation labels to compute eye gaze direction. The left and right eye position levels are the antecedents of the fuzzy system, and they are all divided into three folds: low, medium, and high. The universe of discourse of each eye is dependent on the coordinates of the inner corners and the outer corners of the respective eye; thus, the fuzzy system is locally calibrated to accommodate different eye shapes and sizes. Both iris locations are subsequently fed into the fuzzy system as input for inference.

Attention Assessment Membership Function

The gaze direction variable also employs the Gaussian membership function in the form of equation 3.1. As mentioned, the universe of discourse of this consequent variable is static. Therefore, the means of the three Gaussian functions to describe

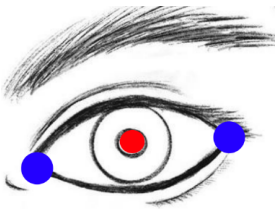


Figure 3.10: Eye key points for gaze assessment.

the three terms in the fuzzy set are 0, 0.5 and 1, with a common standard deviation of 0.1. Graphically, the membership functions are as follows;

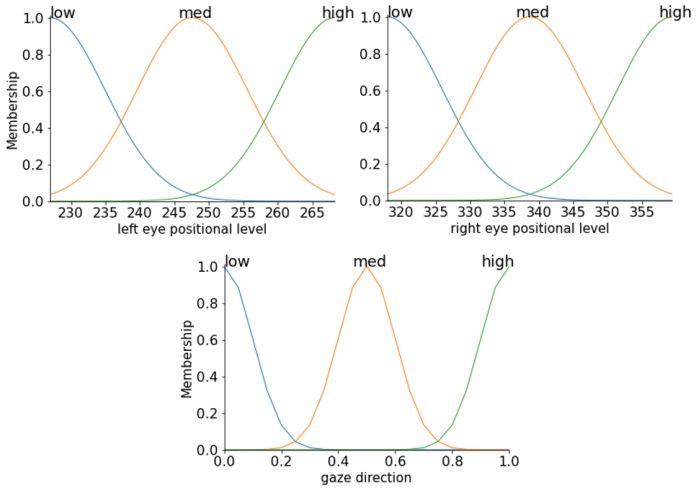


Figure 3.11: Eye gaze membership functions, colors in respective membership functions; top panels show input fuzzy membership functions (Antecedents) and the bottom panel shows the output fuzzy membership function (Consequent).

Attention Assessment Knowledge Base

If there is a need for clarification, the robot needs to understand whether the human partner is paying attention to the robot before it initiates any disambiguation processes; thus, we need to acquire information regarding the gaze contact of the human partner. The human partner is likely not paying adequate attention to the robot if the user is not looking at the robot, and vice versa. The inference rules for the gaze contact remain highly intuitive and can be summarized as follows:

1. **IF** (*left eye position level* is low) **OR** (*right eye position level* is low) **THEN** *overall gaze direction* is left

2. **IF** (*left eye position level* is medium) **OR** (*right eye position level* is medium) **THEN** *overall gaze direction* is center

3. **IF** (*left eye position level* is high) **OR** (*right eye position level* is high) **THEN** *overall gaze direction* is right

The crisp value is again calculated with the centroid method [1]. The crisp value is then passed into thresholding functions to enable the robot to be aware of human eye contact or acknowledge that the user is looking elsewhere.

After we establish the human gaze focus label, we combine it with the human head orientation label to estimate human attention:

1. **IF** (*head orientation* is center) **AND** (*gaze focus* is eye contact) **THEN** *attention* is full attention

2. **IF** (*head orientation* is center) **AND** (*gaze focus* is eye away) **THEN** *attention* is semi-attention

3. **IF** (*head orientation* is left **OR** right) **AND** (*gaze focus* is eye contact) **THEN** *attention* is semi-attention

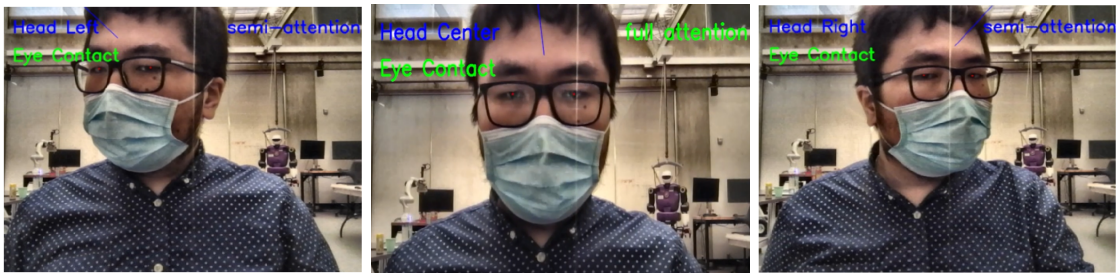
4. **IF** (*head orientation* is left **OR** right) **AND** (*gaze focus* is eye away) **THEN** *attention* is no attention

5. **IF** (*head orientation* is down) **THEN** *attention* is no attention

6. **IF** (*head orientation* is up) **THEN** *attention* is no attention

Example Output

The proposed attention assessment system can determine human attention by head orientation with key points analysis and gaze focus detection with fuzzy inference. The system can robustly evaluate human attention even with obstructions caused by users wearing a face mask or eyeglasses due to our selection of facial key points, and the system is able to determine whether the user is looking at the robot even when the user's head is tilted.



(a) Semi-attentive (b) Fully-attentive (c) Semi-attentive

Figure 3.12: Attention estimation with heavy facial obstruction and different head postures. (a) Semi-attentive (head tilted to the left, eye gaze at the robot); (b) Fully-attentive(head and eye gaze directed at the robot); (c) Semi-attentive (head tilted to the right, eye gaze at the robot)

3.1.5 Disambiguation

When the robot detects that there is ambiguity in the task, and after ensuring that it has adequate human attention, the robot verbally informs the user about the ambiguity. The robot uses a deictic gesture (the robot closes its gripper to form a pointing finger) to point at a candidate object and verbally inquires for feedback. At the same time, the robot alternates its gaze between the object in question and the user to establish gaze communication. The framework estimates human head gestures by employing the Caffe model [20] to detect and locate the face, and it exploits the Lucas-Kanade optical flow method to track the motion of the head [28]. Furthermore, users can provide verbal feedback to the robot. We employ a multi-layer, bi-directional LSTM recurrent neural network trained with the Twitter US Airline Sentiment dataset [21] to classify speech feedback. The feedback network is illustrated as part of Figure 3.14. If the robot receives negative feedback on the selected item, the robot points to the next most confident potential target candidate (confidence scores by YOLO) to continue the disambiguation process. After either the user gesture or the speech feedback is analyzed and the correct reference item is clarified, the ambiguity is eliminated and the robot proceeds to execute the task and engage in learning, if necessary.

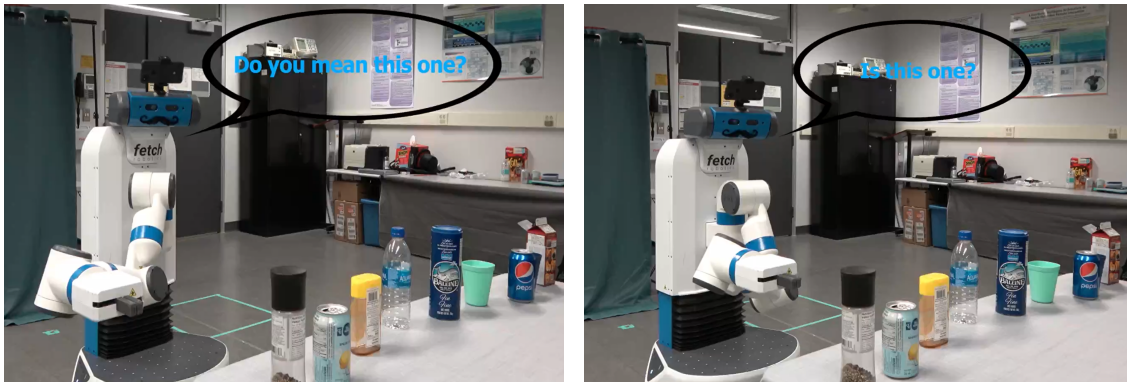


Figure 3.13: Robot disambiguation with speech and gestures.

3.1.6 Learning and Object Selection

When the framework recognizes that the label of the object is novel, it is important to learn the object, so that when the same object is requested again, the robot can proceed without ambiguity. The YOLO network trained with COCO is satisfactory at detecting common objects. However, the network's knowledge of the object is often not granular enough. Though the YOLO model can detect different shapes, textures, and colors, for generalization purposes and because of the COCO dataset's labels, many visually distinct objects are classified into the same class. If the user asks for a "blue cup", the robot will run into task-halting confusion due to a lack of knowledge, as it recognizes all cups with the generic "cups" label. It is, therefore, important to have the means to update the robot's knowledge base through interactions.

The framework employs a short-term/long-term memory scheme to facilitate learning. When an unfamiliar object label is supplied and the robot has no prior experience with the object label, the robot looks at all available pickable objects. The framework stores all images of pickable objects in the short-term memory along with other information (names of objects, base-frame/map-frame coordinates of objects). The short-term memory is volatile, so the memory elements will be destroyed after usage. After all target candidates are stored, the robot disambiguates through social referencing as described in subsection 3.1.5. Once the correct reference is clarified, the referenced object's short-term memory item is transferred to the persistent long-term memory for later recall. The robot approaches the target, turns toward the target object and picks up the appropriate object. Note that the learning process described only occurs when the framework determines that the object label is unfamiliar. It would be inefficient to engage in learning if the base object detection

model already knows the object.

When the robot is asked for an item that has been learned previously, the framework extracts all pickable items seen by the robot's camera and transfers them into the short-term memory, and it invokes the robot's relevant long-term memory element. Short-term memory images are compared against the long-term memory exemplar with a custom siamese network for one-shot learning [22]. The correct item in the short-term memory should have the highest similarity score with the exemplar in the long-term memory and can be retrieved easily. The siamese network we implemented employs CNN twin networks in the well-tested VGG19 architecture [44] and we trained our network with the empirically chosen (among mean squared error, cross-entropy loss and contrastive loss) contrastive loss described below:

$$L = mean((1 - Y) * (Y_{pred})^2 + Y * (max(M - Y_{pred}, 0)^2)), \tag{3.6}$$

where M is the margin, empirically set to 1, the siamese neural network computes the similarity score (Y_{pred}) between the query input and the anchor input. We empirically tested different distance metrics such as the L1, L2 distance and the cosine similarity, and we determined that the L1 distance of the feature embedding perform most satisfactorily in the given task:

$$L1 = \sum_{i=1}^{n} |anchor_i - query_i|. \tag{3.7}$$

The siamese network works in parallel with the YOLO model as an auxiliary; thus, the original objects from YOLO are not modified, and major issues like catastrophic forgetting in online learning are successfully prevented. The learning network is represented as part of Figure 3.14.

Till now, we have explained all the fundamental functional components of the base social referencing disambiguation robot framework. The base framework enables a service robot to analyze various signals and form bidirectional interactions with the human user to disambiguate and learn in object selection tasks. In the next section, we elaborate on the extended framework, where the framework is better at handling novel objects through the robot mental imagery/imagination process, and the overall interaction is more robust via the addition of passive user objection detection.

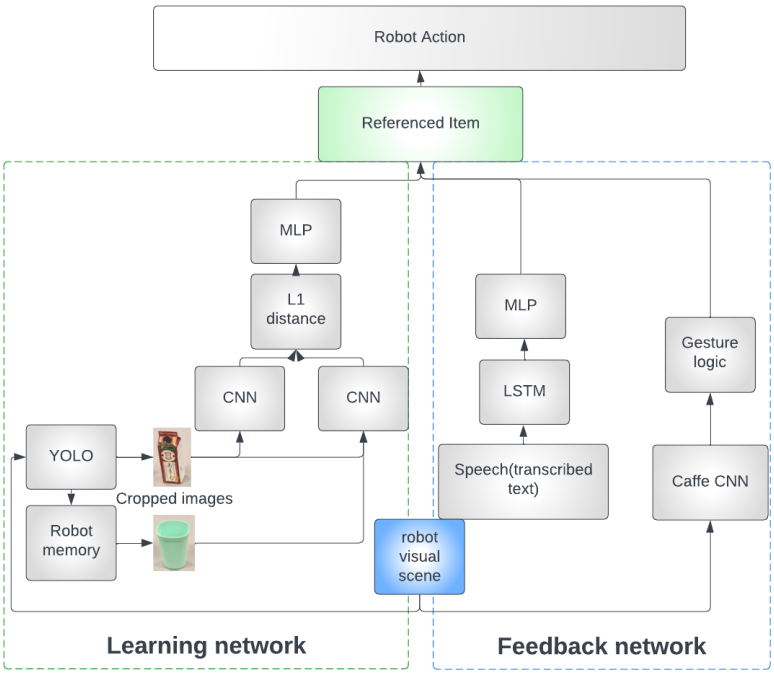


Figure 3.14: Network overview of the social referencing disambiguation framework. Learning network with robot imagination (left), feedback network with speech analysis and head gesture analysis (right).

3.2 Extended Framework

Let us revisit the scenario depicted in Figure 3.2, a human user is having a meal on one side of the table and instructs the service robot to "pass the salt". The robot's base knowledge (its base object detection model) is not sufficient to locate the correct container, as the robot's base knowledge is not granular enough, i.e. the robot is trained to recognize generic containers, but it has no knowledge of the content or the utility of the object. The robot can utilize our proposed social referencing disambiguation framework to locate the correct object and learn the object through interactions with the human user. However, in the base framework, disambiguation can be slow as the robot has no guidance when it encounters a novel object. In the base framework, if the robot detects that the ambiguity is primarily originated from object novelty, the robot will engage in social referencing disambiguation to enquire about the potential target. If there are multiple potential targets, the robot

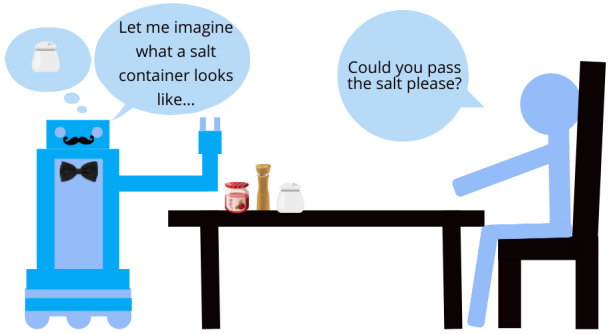


Figure 3.15: The robot does not know which container contains "salt" due to its limited knowledge base; the robot invokes imagination and learning and performs disambiguation with human feedback if the target is still unclear after imagination; Robot's imagination can provide mental guidance to accelerate the disambiguation and learning process.

will naively inquire about objects in an order based on their detection confidence levels. This inquiry order is not optimized, as the confidence level of the object does not correspond to the actual context of the target item. We want a service robot to utilize our proposed social referencing framework to disambiguate the target reference naturally with human feedback and update its knowledge base continuously, through interaction, to resolve ambiguous states in the future. We also need the robot to have the ability to generate visual guidance of the target object (robot mental imagination, visual images of targets) to enhance the robot's disambiguation and learning process. A conceptual illustration of the extended framework can be seen in Figure 3.15. In addition to robot's visual imagination ability, we introduce a second loop of social referencing to enhance the robustness of the framework which will be explained later in this section. The extended framework architecture is illustrated in Figure 3.16.

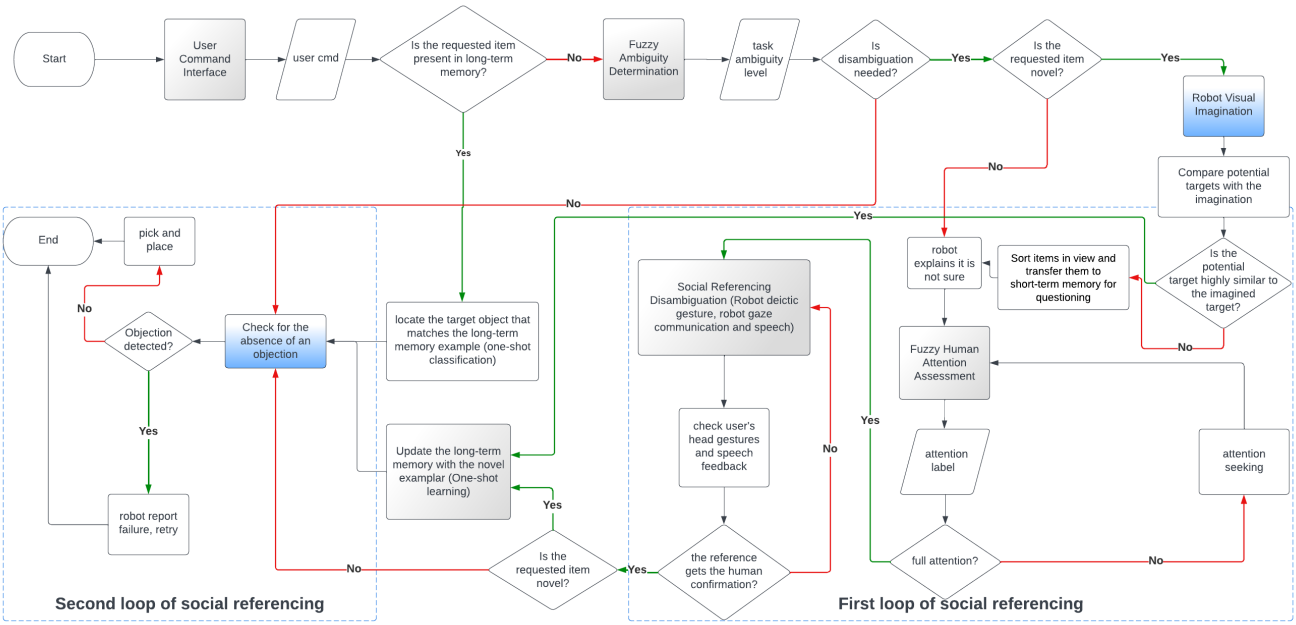


Figure 3.16: The high-level flow diagram of the extended social referencing disambiguation framework with robot mental imagery and a second loop of social referencing.

3.2.1 Robot Mental Imagery

As described in section 3.1.6, a short-term/long-term memory scheme works in parallel with the robot's base knowledge schema. If the robot determines the target label is novel, relevant data such as the object's name, visual appearance (in the format of images), and map/robot frame coordinates of potential candidates are stored in the short-term memory of the robot. Additionally, the robot would engage in a "mental imagination" process based on the latent diffusion method proposed in [38]. The latent diffusion model with T latents $x_1 ... x_T$ is in the form of [19]:

$$p_\theta(x_0) := \int p_\theta(x_{0...T})\, dx_{1...T}, \tag{3.8}$$

where x_0 is the original image, $p_\theta(x_0)$ is the probability the generative model assigns, and the joint probability $p_\theta(x_{0...T})$ represents the denoising reverse process.

The latent diffusion model denoises noise-filled samples by iteratively removing the normally distributed noise in the reverse process. This reverse process corre-

sponds to the learning of Gaussian transitions of a fixed Markov chain. There is a deterministic forward process that is fixed to a Markov chain to gradually add Gaussian noise to the sample. More details on the training and tuning of the model can be found in [38]. In our system, we configure the unconditional guidance scale to be 18 to make the model follow the user's speech prompt more precisely, and we set the speech prompt of the model input to be "a singular, complete, full front view, $'userinput'$ in the center, grey background, photo-realistic", where $'userinput'$ is the extracted object label from the user's instruction.

Given a novel object label, the robot "imagines" one possible visual appearance of the target item based on the speech instruction of the user. More detailed user descriptions give the latent diffusion model a more precise speech prompt to generate more concrete robot mental visual imagery. Examples of the robot's visual mental imagery can be seen in Figure 3.18. After the formation of the robot's mental visual imagery, the robot runs object detection [40] on the generated image to locate the imagined target (Figure 3.17), effectively, this is analogous to the robot 'paying attention' to the most salient object in an imagined scene, and locating the imagined object provides the custom siamese network with the expected data form (the network expects objects, not entire visual scenes), thus enhancing network performance. In this process, the recognition label is unimportant; only the bounding box is necessary as the robot needs to isolate the most salient object in the imagined scene. However, if the object detection fails in this process, i.e., no bounding box can be formed, the robot would keep the entire imagined visual scene for later analysis instead of the isolated visual object. After the extraction, the robot would compare potential candidates in its short-term memory (what the robot sees) with the imagined target candidate (what the robot thinks) via a custom siamese neural network described in Section 3.1.6.

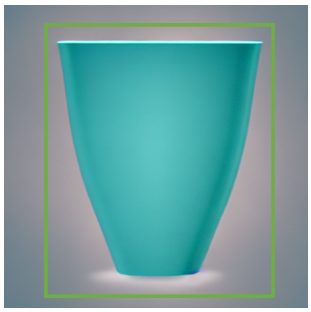


Figure 3.17: Extracted most salient object in the imagined scene for later analysis.



Figure 3.18: Examples of robot's visual mental imagery.

In the case where the robot's generated mental image is highly similar (over 80% similarity) to one of the potential candidates, the ambiguity of the reference selection task is eliminated, and the robot can execute the object fetching and learning without asking any additional questions; however, if the mental image does not closely resemble any of the potential targets in the short-term memory. The memory items will be sorted in descending order based on the memory item's similarity with the mental image. The robot will inquire about the object reference as described in 3.1.5. However, the robot can now disambiguate in a contextual fashion, based on the visual resemblance of the reference and the robot's mental imagery. Once the target reference is clarified, either directly with highly similar robot mental imagery or with social feedback from the user, the associated short-term memory item of the target reference is transferred into the long-term memory, with the updated granular information learned through the interaction. The updated network architecture of the extended framework can be seen in Figure 3.19.

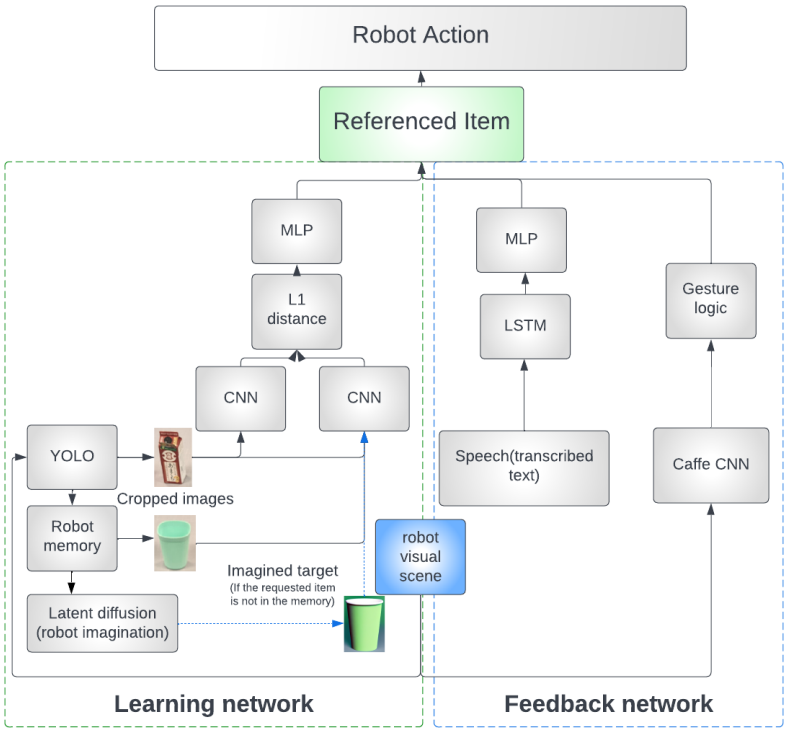


Figure 3.19: Network overview of the extended social referencing disambiguation framework.

3.2.2 Second Loop of Social Referencing

As shown in Figure 3.16, before the robot proceeds to fetch the target item, the robot will be passively observing the human user to seek the 'absence of objection'. If no objection is detected while the robot is preparing to fetch the object, the robot will proceed to execute the task. If an objection from the user (speech or gesture) is detected, then the robot will halt task execution, acknowledge mistakes and retry. This behavior of seeking the *absence of objection* before task execution forms the second loop of social referencing. This simulates the passive, indirect social referencing behavior in humans [13], and can increase the overall robustness of the system. In the human-human interaction domain, this is similar to a person in training trying to assemble an object, selecting a tool based on interaction with the instructor and social referencing, and while grasping the tool, checking with

the instructor 'to make sure' the choice is correct. This second loop can increase robustness in safety or otherwise critical object selection tasks. E.g., in the case of a home companion or health support robot, making the correct selections (e.g., in terms of medications delivered) can be of life-saving importance, while in the context of robotic co-workers (cobots), mistakes made during assembly tasks can have severe financial and safety critical consequences.

3.3 Interaction Example

We now provide a detailed step-by-step description of the results generated by the framework for the case where the robot must handle ambiguity due to object novelty. First, the user asked, "Can you please pass me the green cup?" The robot extracted "green cup" as the target label from the instruction and proceeded to analyze the ambiguity of the task. The robot recognized that the object label was novel; therefore, it could not utilize YOLO to locate the "green cup" among all objects. The robot utilized the stable diffusion model to generate a mental image of the target object (Figure 3.20).The robot scanned the workspace and saved all pickable objects in view in the short-term memory for future disambiguation. As shown in Figure 3.21, the resultant ambiguity is classified as "very high", and the final ambiguity level is as high as 0.918.



Figure 3.20: Robot mental imagery of a 'green cup'.

The robot compared its imagined target with all potential targets in its short-term memory with the siamese neural network described in 3.1.6. In this particular run, the robot found no potential objects have a similarity score higher than the direct learning threshold of 80% (if a potential target reaches the the direct learning threshold, the object is considered as highly similar to the robot's mental imagery, then the robot can directly proceed to locate the correct reference without triggering

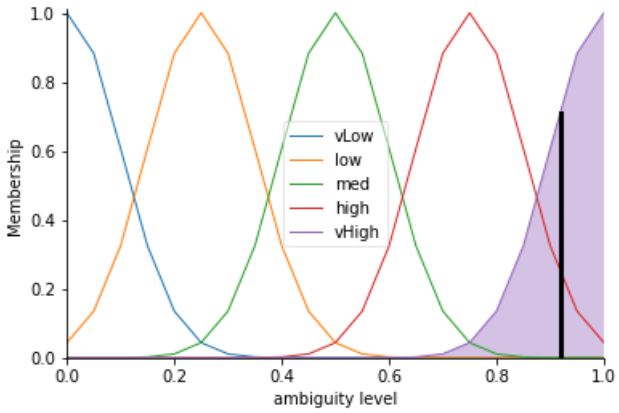


Figure 3.21: Ambiguity level result, ambiguity level can be classified as "vHigh" which means very high, "high" which is high, "med" as medium, "low" as low and "vLow" as very low.

the disambiguation and engage the learning process when appropriate). Though the robot couldn't locate the correct reference directly through mental imagery in this particular execution, the robot could utilize this imagined target item as visual guidance to ask questions effectively. All potential items were compared against with the robot's imagined target, and their corresponding similarity scores were computed. The robot sorted all potential targets in its short-term memory based on the item's visual resemblance with the imagined target in descending order, i.e. the object resembled the robot's mental imagery the most was placed at the first sequence, the the second most similar item and so on. The item that had the highest visual resemblance with the robot's mental image was likely target and the clarification of this object would be inquired first.

The robot then proceeded to assess the user's attention before disambiguation. The robot determined the user was semi-attentive because the user's head was facing the robot, but the user had no eye contact with the robot; thus, the robot verbally requested more attention and the user shifted his gaze toward the robot to provide full attention for the interaction 3.22.

The robot disambiguated through its pointing gesture, gaze communication, and verbal queries as described in 3.1.5. The robot recognized head gestures and it listened to verbal feedback from the user. The user sometimes would say phrases that were irrelevant to the interaction, such as "Nice weather we have today." The language model classified those phrases as irrelevant and ignored them. After clarifying

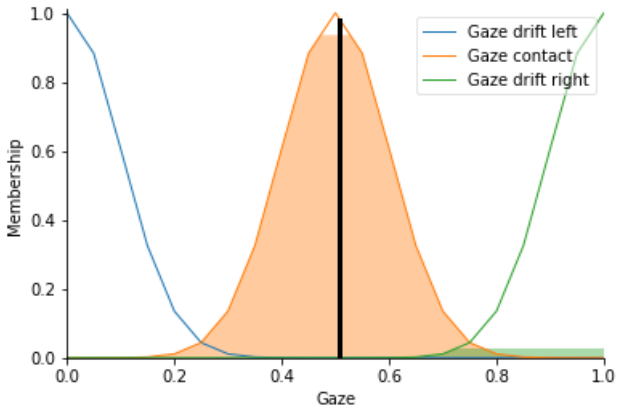


Figure 3.22: Gaze level after attention attraction.

the target item, the robot saved the target item in the long-term memory to associate the new label with the object for future retrieval. When the learned object was requested again, the robot retrieved the relevant memory in the long-term memory and utilized the siamese network to distinguish the target item from all objects in view. Example comparisons can be seen in Figure 3.23.



Figure 3.23: The siamese similarity comparison: the left vertical pair has a similarity score of 0.72, and the right vertical pair has a similarity score of 0.95. The top query images are in the robot's view, and the bottom anchor images are in the robot's long-term memory.

Finally, the robot successfully located the target object based on the highest similarity score with the memorized sample. The robot approached the target item and passively observed the human for the absence of objection (the second loop of

social referencing). The robot did not detect any objection from the human user, so it proceeded to fetch the target item.

3.4 Preliminary Testing

We developed an experiment to test out the base framework. The experiment here provides preliminary verifications of the framework's integration and overall functionality.

The robot and the experimenter are separated by a table. Objects are placed on one side of the table, and a robot running the framework is instructed to pass on a target item by the experimenter. The experiment includes three tests as follows:

- **Test 1**: A set of objects is placed on the table and the target item is unique among all other items. The name of the target item is *known* to the base object detection model. Test 1 is designed to verify if the framework can successfully recognize unambiguous situations and proceed with task execution without the disambiguation process.
- **Test 2**: There are multiple candidates for the potential target, the name of the target item is *known* to the base object detection model. Test 2 is designed to address the ambiguity involved with multiple potential targets and low confidence/misclassification in the base object detection model.
- **Test 3**: There are multiple candidates for the potential target, the name of the target item is *unknown* to the base object detection model. Test 3 examines the learning ability of the framework. Can the framework learn novel object labels after the interaction? Moreover, can the framework successfully retrieve the correct item based on its memory?

The object positions are periodically randomized between trials, and objects are also added/removed between trials for better generalization.

A trial is considered successful if the correct target item is selected by the robot (the success of object pick up is out of the scope of this paper). All tests are repeated 20 times to compute the success rate of the framework. In addition, test 3 consists of pairs of sub-trials for each run. In the first sub-trial, the user needs to supply a novel descriptive object label to the robot, and the robot will disambiguate to locate

the target item and learn the item. In the second sub-trial, the user will provide the same object label as in the previous sub-trial, and the robot should remember the target item and proceed to pick up the item without ambiguity. Therefore, in test 3, success is considered if and only if both sub-trials of the pair are successful.

We performed all tests described on the physical robot (the Fetch mobile manipulator robot3). The success rate of test 1 is 95%, 90% for test 2 and 80% for test 3. The framework is effective based on our preliminary experimental results.

3.5 Discussion

In this chapter, we have detailed the architectural structure of the learning social referencing disambiguation robot framework. We explained the based framework along with all the fundamental functional components in section 3.1. The base framework allows the robot to mimic the social referencing human cognitive mechanism to perform object disambiguation. Multimodal communications between the robot and the human user are established to form effective and natural interactions, and the base framework enables the robot to learn novel objects continuously based on interactions.

Subsequently, we demonstrated further extensions of the framework in section 3.2. The extended framework incorporates all aspects of the base framework, and in addition, the extensions of robot mental imagination and the second loop of social referencing are added to the framework to enhance the learning and disambiguation processes. The robot's mental imagery component utilizes a powerful generative model to empower the robot to imagine the target item when the robot's base object detection knowledge is not sufficient. The mental image can serve as visual guidance to the robot that allows for more efficient query orders or the skipping of the entire disambiguation process when the reference item is highly similar to the imagined target. The second loop of social referencing allows the robot to passively listen to various human signals before task executions to help the robot avoid erroneous actions. This seeking of the absence of human objection can enhance the robustness of the framework further.

Lastly, in section 3.3, we provided a step-by-step interaction scenario for the full framework. We provided the logical flow of the framework interaction, the intermediate states of the framework functional components, and the final outcome of the interaction. The interaction example demonstrated the framework's capability of disambiguation and learning through social referencing, but the interaction we have

shown was only an isolated example. In the next chapter, we will provide details of a full system validation study with human participants to further validate our proposed system.

Chapter 4

Fetch Robot Social Referencing System Validation

This chapter discusses the motivation, experimental design, and implementation of a comprehensive system validation study of the learning social referencing robot framework. Although we tested the framework's functional components individually throughout the framework's development, and we conducted limited testing on the integrated base framework, which is reported in Section 3.4. Nevertheless, the preliminary tests were mainly there to test out the integration of various functional components into a combined framework, and all testing scenarios were conducted and repeated by one human user (the experimenter). Additionally, the extended framework was not tested in the preliminary test. All in all, in order to further validate our proposed system, we were motivated to develop a full human-robot system validation study.

Participants interacted with the Fetch robot (see Section 3) in four scenarios designed to validate components of the framework. They filled out a pre- and post-questionnaire to assess the user's perception of the robot before and after the interaction.

4.1 Interaction Scenarios and Procedure

In order to validate our system, we devised four different interaction scenarios, each serving a specific purpose:

- **Scenario 1**: In this scenario, a group of objects is placed within the workspace, with one particular item being unique (exactly one possible target). The participant's task is to instruct the robot to select the unique item that is *known* to the robot. The objective of this scenario is to assess whether the system is capable of analyzing a clear and unambiguous scene and successfully executing the task.

- **Scenario 2**: Similar to Scenario 1, a set of objects is present in the workspace, but this time there are multiple potential targets. The participant's instruction remains the same—to guide the robot in selecting a *known* item. This scenario aims to test the system's ability to handle an ambiguous scene with multiple potential candidates, as well as low confidence or misclassification in base object detection. The system must employ a social referencing disambiguation process to accomplish the task.

- **Scenario 3**: Once again, a collection of objects is positioned in the workspace, offering multiple potential targets. However, in this scenario, the participant instructs the robot to choose an item that is *unknown* to the robot. The purpose of this scenario is to evaluate whether the system can effectively analyze an ambiguous scene, interpret novel object instructions, and utilize robot mental imagination, learning, and social referencing disambiguation as needed.

- **Scenario 4**: Building upon the previous scenario, the items from Scenario 3 are rearranged randomly. The participant's task is to instruct the robot to select the item mentioned in Scenario 3. This scenario aims to assess the learning outcome of the interaction. We examine whether the robot can acquire knowledge of novel object labels through interaction, retrieve the appropriate memory item, and locate the target reference without requiring further disambiguation.

To enhance generalization and assess the system's resilience, we regularly introduced random variations in object positions and added or removed objects during each trial. Our focus was on evaluating the system's ability to select the correct item rather than solely on successful object pickup, as stated in this paper. The experimental configuration can be observed in Figure 4.1. Throughout the experiment, an experimenter was present in the room to monitor the framework's outputs and document the outcome of each scenario interaction as failure (0), success (1), or task aborted because of passive objection detection (0.5). The robot operated autonomously without any intervention or logical decision-making interference from

the experimenter. However, in case of emergencies, the experimenter had a stop button to remotely shut down the robot. Participants provided informed consent and completed a pre-questionnaire before engaging with all four scenarios in order. Afterward, they filled out a post-questionnaire. Each individual study session had an average duration of approximately 40 minutes.



Figure 4.1: Experiment setup and the Fetch Robot.

4.2 Participants

Our study involved the recruitment of 45 individuals, consisting of 17 males and 28 females. The age of the participants varied between 18 and 41 years, with an average age of 23.36 ± 5.18 years. The recruitment was conducted at the main campus of the University of Waterloo, and the study received approval from the University of Waterloo Human Research Ethics Board.

4.3 Questionnaire

The pre-questionnaire is consists of an adapted Technology-Specific Expectations Scale (TSES) [2], the adapted Robotic Social Attributes Scale (ROSAS) on robot

competence [8], the Godspeed questionnaire on perceived intelligence [5], and the adapted Perceived Safety with Emotional State (PSE) [53], plus simple demographic questions on partcipant's age and gender, and general technology/robotic experience questions. The pre-questionnaire helped analyze participants' expectations and familiarity with technologies and robotics. The post-questionnaire was based on the adapted user experience questionnaire (UEQ) [41], the same Godspeed, ROSAS, PSE questions asked in the pre-questionnaire, in addition to some questions on robot predictability, confidence in the robot, and the robot's dependability, plus an open-ended question on points of improvement (*Which abilities would you improve or add to Fetch if you were to employ it in a home service setting?* (open-ended question)). The ROSAS questionnaire and TSES employed a 5-point Likert scale, and the Godspeed and PSE utilized a 5-point semantic differential scale. Additionally, we had 10-point scaled questions regarding the robot's predictability, dependability, and participant confidence in the robot. Detail of all questionaires, scaled are shown as follows:

The pre-questionnaire consists of various scales and questionnaires, including the adapted Technology-Specific Expectations Scale (TSES) [2], the adapted Robotic Social Attributes Scale (ROSAS) focusing on robot competence [8], the Godspeed questionnaire assessing perceived intelligence [5], and the adapted Perceived Safety with Emotional State (PSE) [53]. It also includes demographic questions about participants' age, gender, and their experience with technology and robotics. The purpose of the pre-questionnaire is to analyze participants' expectations and familiarity with technology and robotics.

The post-questionnaire is based on the adapted user experience questionnaire (UEQ) [41], the same Godspeed, ROSAS, and PSE questions as in the pre-questionnaire, and additional questions on robot predictability, confidence in the robot, and the robot's dependability. It also includes an open-ended question asking participants to suggest improvements or additional abilities they would like to see in Fetch if it were employed in a home service setting. The ROSAS questionnaire and TSES utilize a 5-point Likert scale, while the Godspeed and PSE employ a 5-point semantic differential scale. Additionally, there are 10-point scaled questions regarding the robot's predictability, dependability, and the participants' confidence in the robot. Detailed information about all the questionnaires and scales used is provided below.

Question
I think the robot will have superhuman capacities.
I think the robot will be more than a machine.
I think the robot will be able to perceive what I am going to do before I do it.
I think I will be able to interact with the robot.
I think the robot will be similar to the robots I see in movies.
I think the robot will understand my emotions.
I think the robot will be able to recognize when I look at it or when I shift my gaze to something else.
I think the robot will have sense of humour.
I think the robot will be able to understand me. I think the robot will be able to read my thoughts.

Table 4.1: Adapted Technology-Specific Expectations Scale (TSES).

Question
How many robots have you interacted with?
How many robots have you built/programmed?
How many robots have you controlled?
Have you interacted with the Fetch robot before?

Table 4.2: General Technology experience questions

What is your impression of the Fetch robot?

Incompetent	○ ○ ○ ○ ○	Competent
Ignorant	○ ○ ○ ○ ○	Knowledgeable
Irresponsible	○ ○ ○ ○ ○	Responsible
Unintelligent	○ ○ ○ ○ ○	Intelligent
Foolish	○ ○ ○ ○ ○	Sensible

Table 4.3: Godspeed questionnaire on perceived intelligence with the 5-point semantic differential scale.

Please rate the following items:

You consider that you have a good level of knowledge about robots.	○ ○ ○ ○ ○	You consider that you have a very poor level of knowledge about robots.
You consider yourself very well-informed about robots.	○ ○ ○ ○ ○	You consider yourself very uninformed about robots.
You consider yourself familiar with robots.	○ ○ ○ ○ ○	You consider yourself very unfamiliar with robots.

Table 4.4: Robotic experience questions.

Do you think a robot (or the Fetch robot) is:
Reliable?
Interactive?
Responsive?
Capable?

Table 4.5: Adapted Robotic Social Attributes Scale (ROSAS) on robot competence.

Please rate on these scales how you felt BEFORE/AFTER the interaction:

Anxious	○ ○ ○ ○ ○	Relaxed
Agitated	○ ○ ○ ○ ○	Calm
Quiescent	○ ○ ○ ○ ○	Surprised

Table 4.6: Adapted Perceived Safety with Emotional State (PSE).

Was Fetch

Obstructive	○ ○ ○ ○ ○	Supportive
Complicated	○ ○ ○ ○ ○	Easy
Inefficient	○ ○ ○ ○ ○	Efficient
Confusing	○ ○ ○ ○ ○	Clear
Boring	○ ○ ○ ○ ○	Exciting
Not interesting	○ ○ ○ ○ ○	Interesting
Conventional	○ ○ ○ ○ ○	Inventive
Usual	○ ○ ○ ○ ○	Leading edge

Table 4.7: Adapted User experience questionaire (UEQ)

Questions
To what extent can the robot's behavior be predicted from moment to moment?
What degree of confidence do you have that the robot will be able to cope with similar situations in the future?
To what extent can you count on the robot to do its job?

Table 4.8: Robot predictability, human confidence, and robot dependability

1	2	3	4	5
Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree

Table 4.9: 5-point Likert scale.

label	negative	↔	positive
degree of measurement	1	2 3 4 5 6 7 8 9	10

Table 4.10: 10-point scale of measurement degree.

4.4 Experimental Results

System Performance

We conducted objective measurements to evaluate the outcome of each interaction, as described in Section 4.1. The success rate (in the scale of 0 to 1) of the framework in each interaction was computed as follows: scenario 1: 0.92 ± 0.26, scenario 2: 0.84 ± 0.37, scenario 3: 0.89 ± 0.24, and scenario 4: 0.9 ± 0.27. These scenarios assess different abilities of the framework, and the success rate provides an objective assessment of its performance in handling various types of uncertainties object selection tasks. It is worth noting that scenario 2 exhibited the lowest success rate, which was expected as it involved more interactions with the user. We discovered that object detection is sensitive to factors such as lighting conditions, image quality, and view angles. Enhancing the base object detection model can improve the framework's performance across all aspects. Additionally, we identified that faulty speech transcription can disrupt the interaction's fluency. Improving speech recognition can enhance the framework's interactivity and promote smoother interactions. Furthermore, we observed that the robot's overall performance could be affected by lower battery levels, especially during consecutive participant sessions. Though we understand that the computation heavy sides of the framework are handled by the remote computer, the performance drop when the robot battery level is low is fairly apparent. This performance decline may be attributed to the battery management system's energy conservation efforts, thus causing the robot's sensor to give delayed or inaccurate readings to the framework to analyze. To mitigate this issue, maintaining the robot in a near-fully or fully charged state is recommended, although it may be less relevant or absent on different robotic platforms.

Perceived Robot Attributes

To ensure the reliability and consistency of our data, we conducted Cronbach's alpha tests. The results showed that our pre-questionnaire data had a Cronbach's alpha of 0.88, indicating good data consistency. Similarly, our post-questionnaire data exhibited a Cronbach's alpha of 0.93, indicating excellent internal data consistency. These findings assure us that the collected data is reliable and consistent, supporting the validity of our study.

Our study included a population that can be characterized as highly educated, with each participant having completed at least one year of post-secondary edu-

cation. However, it is important to note that our population represented diverse academic backgrounds, meaning that not all participants had universal experience or familiarity with technology and robotics. On average, participants had previously interacted with an average of 1.80 ± 2.84 robots, controlled (remote/wired) an average of 1.07 ± 1.88 robots, and built/programmed an average of 0.51 ± 1.10 robots.

To assess participants' expectations related to technology and robotics, we computed the technology/robotic expectation score by averaging responses to all questions in the Technology-Specific Expectations Scale (TSES). The expectation score ranged from 1 to 5, with 5 representing the highest level of expectation. Across our population, the overall average expectation score was 3.10 ± 1.33, indicating a moderate level of expectation toward technology and robotics.

In order to examine potential differences in perceived intelligence, robot competence, and user-reported emotional state before and after the interaction, we conducted a two-sample paired t-test on the utilized Godspeed, ROSAS, and PSE questionnaires. This statistical analysis aimed to determine whether there were significant variations in these measures following the interaction compared to the initial assessments.

Construct	Category	Before	After	P-value
Godspeed	Competent	4.04 ± 0.71	4.20 ± 0.94	0.37
	Knowledgeable	3.73 ± 0.81	4.09 ± 0.92	**0.0079***
	Responsible	3.8 ± 0.82	4.16 ± 0.95	0.056
	Intelligent	3.76 ± 0.98	4.29 ± 0.79	**0.0012***
	Sensible	3.56 ± 0.78	4.13 ± 0.76	**0.00023***
ROSAS	Reliable	3.80 ± 0.79	3.80 ± 1.01	1.00
	Interactive	4.00 ± 0.98	4.40 ± 0.84	**0.030**
	Responsive	4.02 ± 0.78	4.27 ± 1.05	0.19
	Capable	3.98 ± 0.81	4.11 ± 0.91	0.39
PSE	Relaxed	3.75 ± 1.11	4.40 ± 0.84	**0.0020***
	Calm	4.16 ± 0.77	4.42 ± 0.92	0.10
	Engaged	3.33 ± 1.00	4.04 ± 0.90	**3.53e-5***

Table 4.11: In examining participants' responses before and after their interaction with the system, we have reported the mean scores. Significant differences with a p-value less than 0.05 have been highlighted in bold, while items that retained their significance even after applying the Benjamini and Hochberg False Discovery Rate correction have been denoted with an asterisk (*).

After analyzing the data, we found that participants perceived the robot to be significantly more knowledgeable, intelligent, sensible, and interactive following their interaction with the system. Several factors likely contributed to this positive perception, including the robot's ability to learn through social interactions and imagination, its comprehension of human gestures and speech, and its own use of gestures, speech, and gaze during interactions. However, not all categories of perceived intelligence and competence showed significant differences. We speculate that the robot's navigation and object manipulation skills, the speed of its mental imagination process (which averaged 36 seconds per use of imagination), and occasional delays in speech processing and gesture detection due to device networking could negatively influence participants' perception of the robot.

In addition, we asked participants to rate the robot's predictability, their confidence in the robot, and the robot's dependability on a scale of 1 to 10 (with 1 being the lowest and 10 being the highest). The average ratings for predictability, confidence, and dependability were found to be 7.20 ± 1.55, 8.18 ± 1.71, and 7.36 ± 1.79, respectively. The high level of predictability aligns with the observed significant difference in participants' relaxed state after interacting with the robot.

Furthermore, we calculated the average scores of the User Experience Questionnaire (UEQ) following the interaction in order to gain deeper insights into the overall user experience.

Construct	Category	Mean score
UEQ	Supportive	4.29 ± 0.79
	Easy	4.06 ± 0.96
	Efficient	3.67 ± 0.95
	Clear	4.09 ± 1.00
	Exciting	4.40 ± 0.78
	Interesting	4.62 ± 0.61
	Inventive	4.11 ± 0.86
	Leading edge	4.00 ± 0.80

Table 4.12: The user experience following the interaction (5-point scale).

As indicated in Table 4.12, all user experience categories (5-point scale) surpass the neutral 3-point threshold, with an overall user experience score of 4.14 ± 0.28. Notably, the excitement level receives high ratings, which aligns with the significant difference observed in the Engaged state after the interaction (Table 4.11). However, we observe that the efficiency category receives the lowest rating compared to the

other categories. This observation is consistent with the feedback received through open-ended questions, where the most common point mentioned (11 responses of 45) relates to the robot's motor motions and response time. Additionally, enhancing the clarity aspect can be achieved by increasing the robot's speech production, as six open-ended responses suggested providing the robot with more speech capabilities and utilizing robot speech more frequently. The remaining open-ended responses are not pertinent to the system's development, such as improving the physical appearance or mechanical characteristics of the robot. It is evident that many of these issues could be addressed or mitigated with a more agile and / or humanoid robotic platform.

Chapter 5

Conclusion

This thesis presents a robot framework inspired by the human cognitive mechanism of social referencing. The proposed robot framework allows the service robot to assess task scenes and human states and form bidirectional communication (gesture and speech) between the robot and the user. Importantly, the framework enables the robot to gain granular knowledge about objects through natural interactions with the user. Furthermore, the framework enables robots to utilize visual imagination to optimize social disambiguation, and learning processes. To validate the functionality of our system and to understand the user perception of the robot running the framework, we developed a system validation experiment that involved 45 participants interacting with the robot in interaction scenarios that tested out different functional components of the framework. The study results objectively validated our system and provided valuable information about how the framework influenced the user's perception of the robot.

5.1 Summary of Contributions

The contributions of this research can be summarized as follows:

- We developed fuzzy inference-based solutions for robot object selection task ambiguity determination and human attention assessment. Our fuzzy ambiguity determination functional component allows the robot to analyze various sources of ambiguity and understand how ambiguous the task is, so that the robot can proceed to generate appropriate reactions adaptively. The human

attention assessment functional block enables the robot to assess the user's attention through the analysis of head orientation and gaze direction. The gaze direction can be robustly captured through the fuzzy inference engine so that the robot will always have an understanding of the user's gaze fixation on the robot. The fuzzy functional components are modular and modifiable when new domain knowledge can be acquired, and the fuzzy logic more closely reassembles the human's probabilistic decision-making process than traditional Boolean logic.

- A learning social referencing disambiguation robot framework has been developed. The framework integrated multiple functional components to allow the robot to understand task ambiguity, estimate human attention, and disambiguate and learn objects through natural interactions with the user. The framework is human-inspired and explores the practical utilitarian aspects (disambiguation and discovery) of social referencing. It exploited modern deep learning methods to resolve various ambiguities in robot object selection tasks.

- We further expanded the base social referencing framework with human-inspired extensions to enhance the framework's capability and showcase the framework's extendability and modularity. The robot mental imagery extension empowers the robot to generate mental images of target objects, and the robot can utilize the imagined target to guide its disambiguation process or even locate the correct reference directly in cases where the imagined target closely resembles the actual target object. The second social referencing component allows the robot to passively listen for signs of objection from the user before the final task execution. This component provides the user with additional opportunities to correct the robot's actions and improve the overall reliability of the system. This function is particularly valued in safety-related tasks such as hazardous object handling.

- We developed a human-robot system validation experiment to test out our proposed system. We determined that the system functions satisfactorily through objective observations and measures. In addition, we discovered that the user's perception of the robot was significantly impacted in a positive manner after interacting with the robot running the framework.

We will also provide our research question again, and their answers corresponding answers:

RQ1: How can we create a social referencing framework for service robots to handle object selection tasks?

We designed the framework in a modular fashion where the framework is consisted of five fundamental functional components: *User command interface*, *Fuzzy ambiguity determination*, *Fuzzy human attention assessment*, *Social referencing disambiguation*, *Short-term long-term memory object learning*. These five components form the base social referencing framework where input capture, scene understanding, human states estimation, bidirectional object disambiguation and learning through interaction can be achieved. Later the framework is further expanded with *Robot mental imagery/imagination* and *Objection absence detection* where the framework is better at handling novel objects can form more robust interaction compared to the base framework.

RQ2: How can we create a social referencing framework for service robots to handle object selection tasks?

Chapter 3 demonstrated the feasibility of the framework from both a technical and implementation standpoint. We developed and effectively implemented the framework on a physical robot manipulator, Fetch. Following this, in Chapter 4, the system validation experiment validated the framework's feasibility from an application and user acceptance perspective.

RQ3: How can we create a social referencing framework for service robots to handle object selection tasks?

In Chapter 4, the study results revealed that users perceived the robot with the framework in a highly positive manner. Following interactions with the robot operating the proposed framework, users reported perceiving the robot as more *knowledgeable*, *intelligent*, *sensible*, and *interactive*. Furthermore, users felt at ease during their interaction with the robot, finding the experience engaging. The overall user experience score surpassed the neutral point, providing further evidence that the computational framework underlying the robot's behaviour supported users' perceptions of the robot.

All in all, our proposed framework enables service robots to disambiguate objects and learn through human-robot interactions. The market potential and the societal benefits of intelligent service robots are still largely undiscovered/underdeveloped,

and future service robots need to have the ability to interact with humans socially and effectively and be able to constantly adapt to new environments and update beliefs through accommodating new knowledge. This framework provided an important foundation for future service robots to become social "lifelong" learning agents.

5.2 Lessons Learned

The development of frameworks and systems for service robots is an extremely challenging endeavor. For industrial robots, tasks are highly deterministic and can be broken down into fundamental steps to form relatively simple procedures for robots to execute. More importantly, industrial robots need to show little or no consideration toward humans, as industrial tasks are highly goal-oriented. Industrial robots serve the *industry*, their purpose is to achieve higher industrial goals and make work processes more efficient and effective. Fundamentally differing from industrial robots, service robots need to be astronomically more adaptive and intelligent, thus usually more complex due to their distinctly unique purpose: service robots need to serve the *human*, they need to ensure they can achieve assistive tasks that are acceptable to humans, understandable to humans, and desirable to humans. The *human* is a critical part of the service robot's objective equation when performing any task, and because human behaviors can be extremely sophisticated, dynamic, and situational, therefore, service robot tasks can be highly uncertain. Understanding human states and intent and effective communication with humans are thus critical abilities that successful service robots need to possess. In addition, the human environment is constantly changing; no designer can foresee all possible scenarios that the service robot may encounter; therefore, it is paramount for service robots to be able to continuously learn from mistakes, experiences, and interactions. The capability of "lifelong" learning is essential for the longevity of human robot interaction, and the practical application value of service robots.

Despite all the difficulties, the value that intelligent service robots can bring to human society is possibly huge. Service robots could positively impact domestic, educational, hospitality, and almost all other industries that can provide direct services to humans. Intelligent service robots can become a natural part of people's daily lives and play an important role in human society by liberating humans to focus on creative tasks and activities that humans deem enjoyable. Furthermore, intelligent service robots might be the solution, or at least the partial solution, to the severe labor shortage issue in many advanced but aging societies.

Through this research, we derived a tremendous amount of inspiration from human cognition and development. We reinforced the concept that science is a process of observation and gaining insights from such observations and data. Robots and living creatures share a lot of similarities because we constantly try to make robots to simulate what life or aspects of living can do. We think that bio-inspired robotics is the clear future direction for achieving the general intelligent service robots we envision. Lives are pieces of engineering miracles that are incredibly simple and complex, fragile and robust, stubborn and adaptable. Life is a contradicting manifestation of wisdom and elegance, and nature is man's greatest teacher.

5.3 Limitations

The framework's language understanding and formation, gesture recognition and production functions can be improved for richer, more natural interactions. However, the framework is designed to be modular and extendable, so that large language modals and more sophisticated pose estimation and motion production methods can be integrated.

Currently, if the robot detects user objections before task execution, the robot will halt task execution and retry the task. There is the possibility that the robot may get stuck in a loop of retries, though the retry loop never occurred in our validation study with 45 participants. Appropriate knowledge update mechanisms should be incorporated in future to avoid this logic loop altogether.

This framework is a high-level conceptual prototype for service robots, and it currently addresses the practical issue of robot object selection ambiguity through human-assisted robot disambiguation and learning. However, due to time and resource constraints, the framework is focused on this single-step task (object selection task). The framework is not adapted to multi-step tasks at its current stage, and more time and energy should be invested to make the framework generalize to multi-step tasks.

Our system validation study 4.4 demonstrated the framework's effectiveness and improved human perception on the robot. However, our study only showed the short-term success of the robot framework. After the framework is adapted to more general multi-step tasks, it would be interesting to develop a long-term study to study the long-term effect of the framework.

5.4 Future Directions

There are many exciting future directions for our social referencing learning framework. In the short term, we can improve the current learning paradigm by introducing an extension, so that the robot can better organize its memory items. We envision this as the "contextual memory organization" functional block. With this functional component, the robot should be able to cluster relevant memory items together based on some contexts (utility, appearance similarity, etc.). For example, if the robot clusters based on object appearance similarity, "blue plastic cup" and "blue cup" can be organized together, and if memory items are organized based on utility, then "apple" and "cookies" should be grouped together. There should also be a decision logical block to help the robot decide which memory organization is most appropriate based on the task description.

Related to memory items, we want to compose a knowledge update strategy if the robot detects user objections before task executions. One approach is to make the framework destroy the associated long-term memory item that caused the user's objection so that the robot can relearn the object in the retry process. This approach should help the robot to avoid the possible logic loop mentioned in the limitations section.

Later on, it would also be interesting to incorporate the emotional aspect of social referencing, similar to [6, 45]. We think it might be beneficial for the robot to learn some kind of preference for objects by studying the emotional display of the user. This can allow the robot to have internal rankings of objects to help the robot become more personalized for the user. However, we do need to put more consideration into how to detect human emotion accurately for this extension. Emotion recognition based on facial expression can be inaccurate and unreliable, particularly for weak emotions, and some methods based on physiological signals can be too intrusive for our application [9]. Perhaps we need to pursue a more multi-modal approach (i.e., perform facial, tone, non-invasive, or minimally intrusive physiological signal analysis) to achieve the desired outcome.

Recent advances in large language models (LLMs) like chat-GPT [14] also bring exciting opportunities to extend our framework. It is obvious that we can utilize the language understanding and production capabilities of the LLM to enrich and create smoother and more natural interactions. More interestingly, we can harness the power of LLM to enable the framework to gain a better contextual understanding of the environment and help the robot model itself. LLM can also be applied to help our framework generalize to multi-step tasks once a high-level contextual understanding

of the scene can be established by the robot. The contextual scene construction can be obtained by the robot's perceptual devices, like the RGB camera and LiDAR sensors, and language communication with the human. LLMs can then suggest high-level steps for multi-step tasks. On the point of multi-step tasks, it would also be interesting to expand on our mental imagination functional block. Specifically, it would be beneficial to modify/extend the robot imagination component so that the robot can imagine a sequence of actions when a command is given. The generated images may be considered sequences of goal states to assist the robot in completing complex, multi-step tasks.

References

[1] JDWarner/scikit-fuzzy: Scikit-Fuzzy version 0.4.2, November 2019.

[2] Patrícia Alves-Oliveira, Tiago Ribeiro, Sofia Petisca, Eugenio Di Tullio, Francisco S Melo, and Ana Paiva. An empathic robotic tutor for school classrooms: Considering expectation and satisfaction of children as end-users. In *Social Robotics: 7th International Conference, ICSR 2015, Paris, France, October 26-30, 2015, Proceedings 7*, pages 21–30. Springer, 2015.

[3] Stylianos Asteriadis, Kostas Karpouzis, and Stefanos Kollias. Robust validation of Visual Focus of Attention using adaptive fusion of head and eye gaze patterns. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 414–421, November 2011.

[4] Albert Bandura. Social cognitive theory of social referencing. In *Social Referencing and the Social Construction of Reality in Infancy*, pages 175–208. Springer, 1992.

[5] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, 1:71–81, 2009.

[6] Sofiane Boucenna, Philippe Gaussier, and Laurence Hafemeister. Development of first social referencing skills: Emotional interaction as a way to regulate robot behavior. *IEEE Transactions on Autonomous Mental Development*, 6(1):42–55, 2013.

[7] Joseph J Campos, Seinenu Thein, and Daniela Owen. A Darwinian legacy to understanding human infancy: Emotional expressions as behavior regulators. *Annals of the New York Academy of Sciences*, 1000(1):110–134, 2003.

[8] Colleen M Carpinella, Alisa B Wyman, Michael A Perez, and Steven J Stroessner. The robotic social attributes scale (rosas) development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*, pages 254–262, 2017.

[9] Andrius Dzedzickis, Artūras Kaklauskas, and Vytautas Bucinskas. Human emotion recognition: Review of sensors and methods. *Sensors*, 20(3):592, 2020.

[10] Ethan Eade. Gauss-Newton / Levenberg-Marquardt optimization. https://www.ethaneade.com/optimization.pdf, March 2013.

[11] Kevin Fan, Melanie Jouaiti, Kerstin Dautenhahn, and Chrystopher L Nehaniv. Fuzzy object ambiguity determination and human attention assessment for domestic service robots. In *2022 9th International Conference on Soft Computing & Machine Intelligence (ISCMI)*, pages 140–145. IEEE, 2022.

[12] Kevin Fan, Melanie Jouaiti, Ali Noormohammadi-As, Chrystopher L. Nehaniv, and Kerstin Dautenhahn. A social referencing disambiguation framework for domestic service robots. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11979–11985, 2023.

[13] Saul Feinman. Social referencing in infancy. *Merrill-Palmer Quarterly (1982-)*, pages 445–470, 1982.

[14] Luciano Floridi and Massimo Chiriatti. GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

[15] Juan Miguel Garcia-Haro, Edwin Daniel Oña, Juan Hernandez-Vicen, Santiago Martinez, and Carlos Balaguer. Service robots in catering applications: A review and future challenges. *Electronics*, 10(1):47, 2020.

[16] A. Haasch, N. Hofemann, J. Fritsch, and G. Sagerer. A multi-modal object attention system for a mobile robot. In *2005 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2712–2717, August 2005. ISSN: 2153-0866.

[17] Péter Halácsy, Andras Kornai, and Csaba Oravecz. HunPos-an open source trigram tagger. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 209–212, 07 2007.

[18] Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, and Jethro Tan. Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3774–3781, May 2018. ISSN: 2577-087X.

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[20] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pages 675–678, 2014.

[21] Kaggle. Twitter us airline sentiment. https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment, Oct 2019.

[22] Gregory Koch, Richard Zemel, Ruslan Salakhutdinov, et al. Siamese neural networks for one-shot image recognition. In *ICML Deep Learning Workshop*, volume 2. Lille, 2015.

[23] Stephen M Kosslyn, William L Thompson, and Giorgio Ganis. *The Case for Mental Imagery*. Oxford University Press, 2006.

[24] Séverin Lemaignan, Fernando Garcia, Alexis Jacq, and Pierre Dillenbourg. From real-time attention assessment to "with-me-ness" in human-robot interaction. In *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 157–164, March 2016. ISSN: 2167-2148.

[25] Songpo Li, Xiaoli Zhang, Fernando J. Kim, Rodrigo Donalisio da Silva, Diedra Gustafson, and Wilson R. Molina. Attention-Aware Robotic Laparoscope Based on Fuzzy Interpretation of Eye-Gaze Patterns. *Journal of Medical Devices*, 9(4):041007, December 2015.

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[27] Yen-Chen Lin, Pete Florence, Andy Zeng, Jonathan T Barron, Yilun Du, Wei-Chiu Ma, Anthony Simeonov, Alberto Rodriguez Garcia, and Phillip Isola. Mira: Mental imagery for robotic affordances. In *6th Annual Conference on Robot Learning*, 2022.

[28] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, volume 2, pages 674–679, 1981.

[29] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. MediaPipe: A Framework for Building Perception Pipelines. *arXiv:1906.08172 [cs]*, June 2019. arXiv: 1906.08172.

[30] Aly Magassouba, Komei Sugiura, and Hisashi Kawai. A multimodal target-source classifier with attention branches to understand ambiguous instructions for fetching daily objects. *IEEE Robotics and Automation Letters*, 5(2):532–539, 2020.

[31] Patrick Mania, Franklin Kenghagho Kenfack, Michael Neumann, and Michael Beetz. Imagination-enabled robot perception. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 936–943. IEEE, 2021.

[32] Bill Marina, Müller Christopher, Kraus Werner, and Bieller Susanne. IFR world robotics 2022. https://ifr.org/downloads/press2018/2022_WR_extended_version.pdf, October 2022. Accessed: 2023-5-3.

[33] Jesus Minguillon, M. Angel Lopez-Gordo, and Francisco Pelayo. Detection of attention in multi-talker scenarios: A fuzzy approach. *Expert Systems with Applications*, 64:261–268, December 2016.

[34] Joel Pearson. The human imagination: the cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience*, 20(10):624–634, 2019.

[35] Pradip Pramanick, Chayan Sarkar, Snehasis Banerjee, and Brojeshwar Bhowmick. Talk-to-resolve: Combining scene understanding and spatial dialogue to resolve granular task ambiguity for a collocated robot. *Robotics and Autonomous Systems*, 155:104183, 2022.

[36] Pradip Pramanick, Chayan Sarkar, Sayan Paul, Ruddra dev Roychoudhury, and Brojeshwar Bhowmick. Doro: Disambiguation of referred object for embodied agents. *IEEE Robotics and Automation Letters*, 7(4):10826–10833, 2022.

[37] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. *arXiv:1804.02767 [cs]*, April 2018. arXiv: 1804.02767.

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[39] Deb Roy, Kai-Yuh Hsiao, and Nikolaos Mavridis. Mental imagery for a conversational robot. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(3):1374–1383, 2004.

[40] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[41] Martin Schrepp, Jorg Thomaschewski, and Andreas Hinderks. Construction of a benchmark for the user experience questionnaire (UEQ). 2017.

[42] Mohit Shridhar and David Hsu. Interactive visual grounding of referring expressions for human-robot interaction. *arXiv preprint arXiv:1806.03831*, 2018.

[43] Elena Sibirtseva, Dimosthenis Kontogiorgos, Olov Nykvist, Hakan Karaoguz, Iolanda Leite, Joakim Gustafson, and Danica Kragic. A Comparison of Visualisation Methods for Disambiguating Verbal Requests in Human-Robot Interaction. In *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 43–50, August 2018. ISSN: 1944-9437.

[44] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[45] Andrea Lockerd Thomaz, Matt Berlin, and Cynthia Breazeal. An embodied computational model of social referencing. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, pages 591–598. IEEE, 2005.

[46] Akira Utsumi, Nobuji Tetsutani, and Seiji Igi. View-based detection of 3-d interaction between hands and real objects. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 4, pages 961–964. IEEE, 2004.

[47] David Whitney, Eric Rosen, James MacGlashan, Lawson LS Wong, and Stefanie Tellex. Reducing errors in object-fetching interactions through social feedback. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1006–1013. IEEE, 2017.

[48] Jochen Wirtz, Paul G. Patterson, Werner H. Kunz, Thorsten Gruber, Vinh Nhat Lu, Stefanie Paluch, and Antje Martins. Brave new world: service robots in the frontline. *Journal of Service Management*, 29(5):907–931, January 2018.

[49] L.A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, June 1965.

[50] L.A. Zadeh. Fuzzy logic. *Computer*, 21(4):83–93, April 1988.

[51] L.A. Zadeh. Fuzzy logic = computing with words. *IEEE Transactions on Fuzzy Systems*, 4(2):103–111, May 1996.

[52] Hanbo Zhang, Yunfan Lu, Cunjun Yu, David Hsu, Xuguang La, and Nanning Zheng. Invigorate: Interactive visual grounding and grasping in clutter. *arXiv preprint arXiv:2108.11092*, 2021.

[53] Susana Zoghbi, Elizabeth Croft, Dana Kulić, and Mike Van der Loos. Evaluation of affective state estimations using an on-line reporting device during human-robot interactions. In *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3742–3749. IEEE, 2009.

APPENDICES

Appendix A

System Validation Ethics Approval

UNIVERSITY OF WATERLOO

Notification of Ethics Clearance to Conduct Research with Human Participants

---

Principal Investigator: Kerstin Dautenhahn (Electrical and Computer Engineering)

Student investigator: Kevin Fan (Electrical and Computer Engineering)

Faculty supervisor: Chrystopher Nehaniv (Systems Design Engineering)

Student investigator: Ali Noormohammadi Asl (Systems Design Engineering)

Collaborator: Melanie Jouaiti (Imperial College London)

File #: 44437

Title: Social Referencing Disambiguation for Domestic Service Robots

---

The Human Research Ethics Board is pleased to inform you this study has been reviewed and given ethics clearance.

Initial Approval Date: 10/25/22 (m/d/y)

University of Waterloo Research Ethics Boards are composed in accordance with, and carry out their functions and operate in a manner consistent with, the institution's guidelines for research with human participants, the Tri-Council Policy Statement for the Ethical Conduct for Research Involving Humans (TCPS, 2nd edition), International Conference on Harmonization: Good Clinical Practice (ICH-GCP), the Ontario Personal Health Information Protection Act (PHIPA), the applicable laws and regulations of the province of Ontario. Both Boards are registered with the U.S. Department of Health and Human Services under the Federal Wide Assurance, FWA00021410, and IRB registration number IRB00002419 (HREB) and IRB00007409 (CREB).

This study is to be conducted in accordance with the submitted application and the most recently approved versions of all supporting materials.

Expiry Date: 10/26/23 (m/d/y)

Multi-year research must be renewed at least once every 12 months unless a more frequent review has otherwise been specified. Studies will only be renewed if the renewal report is received and approved before the expiry date. Failure to submit renewal reports will result in the investigators being notified ethics clearance has been suspended and Research Finance being notified the ethics clearance is no longer valid.

Level of review: Delegated Review

Signed on behalf of the Human Research Ethics Board

Heather Dekker, Ethics Advisor, hdekker@uwaterloo.ca, 519-888-4567, ext. 41506

This above named study is to be conducted in accordance with the submitted application and the most recently approved versions of all supporting materials.

Documents reviewed and received ethics clearance for use in the study and/or received for information:

file: FetchRobot_Standard_Operating_Procedure_V1_20220812.docx

file: FetchRobotics_version1_20220518.pdf

file: Consent Form_V3_20221024.docx

file: info_for_research_V3_20221024.docx

file: feedback_letter_uni_V2_20221020.docx

file: setup.pdf

file: pre study Survey.pdf

file: post study Survey.pdf

file: poster_V2_20221013.png

file: email_V2_20221013.docx

Approved Protocol Version 3 in Research Ethics System

This is an official document. Retain for your files.

You are responsible for obtaining any additional institutional approvals that might be required to complete this study.