THE BENEFITS OF OTHER-ORIENTED ROBOT DECEPTION IN HUMAN-ROBOT INTERACTION

A Thesis
Presented to
The Academic Faculty

by

Jaeeun Shim

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Electrical and Computer Engineering

Georgia Institute of Technology
May 2017

Copyright © 2017 by Jaeeun Shim

THE BENEFITS OF OTHER-ORIENTED ROBOT DECEPTION IN HUMAN-ROBOT INTERACTION

Approved by:

Professor Ronald C. Arkin, Advisor
School of Interactive Computing
Georgia Institute of Technology

Dr. Thomas R. Collins
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Ayanna M. Howard
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Sonia Chernova
School of Interactive Computing
Georgia Institute of Technology

Professor Magnus Egerstedt
School of Electrical and Computer Engineering
Georgia Institute of Technology

Professor Alan R. Wagner
Department of Aerospace Engineering
Pennsylvania State University

Date Approved: March 27, 2017

To my husband and family,

ACKNOWLEDGEMENTS

First of all, I would like to thank my advisor, Dr. Ronald C. Arkin, for his insightful guidance and endless support throughout my Ph.D. program. He has been patient and supportive throughout my entire journey, and none of this research could have been possible without his vision and direction.

I also thank my dissertation and proposal committee members, Dr. Ayanna M. Howard, Dr. Magnus Egerstedt, Dr. Thomas R. Collins, Dr. Sonia Chernova, and Dr. Alan R. Wagner, for their priceless academic advice and support for my research. I sincerely thank Dr. Linda Tickle-Degnen, Dr. Andrea L. Thomaz, and Dr. Matthias Scheutz for giving me insightful academic advice, and I appreciate all the help that I received from Mrs. Michaelle Arkin, Dr. Wendy Rogers, and the Human Factors and Aging Lab at Georgia Tech while running my research study.

I want to extend my gratitude to my colleagues in the Mobile Robot Lab at Georgia Tech, Michael Pettinati, Shu Jiang, Matthew O'Brien, Lakshmi Velayudhan for their help and the invaluable discussions about robotics research and life as a graduate student. I am also thankful for the strong support from many of my friends, Yujung Lee, Hayang Kim, Yejin Kim, Minyeong Kim, Sunghee Cho, Minjung Kim, along with all my friends in bay area.

I am extremely grateful to my loving husband, Hanseung Lee, who has supported me and my life as a Ph.D. student with endless love, trust, and understanding. I also want to express my appreciation and respect to my parents, Iksup Shim and Namhi Yoo, for raising me with unending love, and my parents-in-law, Heeyeon Lee and Soyoun Park, for giving me endless support. I give thanks to my sister Jaejin Shim, my brother Jaewoong Shim, and my furry companion Ari for always encouraging me,

and I appreciate all the support from my brother-in-law and his wife, Hansang Lee and Jung Lee. All my other family members and friends also deserve my appreciation, especially my aunt Guisup Shim and twin cousins, Jooyoung and Jeeyoung Lee, in Atlanta who have supported me so much during my PhD studies.

TABLE OF CONTENTS

DEDICATION . . . . . **iii**

ACKNOWLEDGEMENTS . . . . . **iv**

LIST OF TABLES . . . . . **ix**

LIST OF FIGURES . . . . . **xi**

SUMMARY . . . . . **xiv**

I INTRODUCTION **1**

1.1 Research Question 3

1.1.1 Primary Research Question 3

1.1.2 Subsidiary Questions 4

1.2 Objectives 5

1.3 Dissertation Outline 6

II RELATED WORK **7**

2.1 Animal Deception 7

2.2 Human Deception 9

2.3 Robot Deception 16

2.4 Ethical Theory 21

III A TAXONOMY OF ROBOT DECEPTION **24**

3.1 Taxonomies of Deception from a Human Perspective 24

3.2 A Taxonomy of Robot Deception 29

3.3 Robot Deception: A case study 32

3.3.1 Biological Findings 33

3.3.2 Computational Model and Implementation 33

3.3.3 Experimental Results 39

3.3.4 Robot Deception Taxonomy 42

3.4 Summary 43

IV COMPUTATIONAL ARCHITECTURE 44

4.1 Computational model for a robot's other-oriented deception 45

4.2 Method: Deceptive Action Generation Mechanism 46

4.2.1 Deception Generation Mechanism 47

4.2.2 Generating Deceptive Action 47

4.3 Deceptive Action Selection Mechanism 65

4.3.1 Deceptive Action Selection via CBR 69

4.3.2 Exemplar Scenario 89

4.4 Summary 103

V EVALUATING ROBOT DECEPTION IN HRI STUDIES 104

5.1 Potential other-oriented robot deception contexts and Selected HRI study domain 105

5.2 Study Design 107

5.2.1 Study Domain 108

5.2.2 2 by 2 Mixed-subject design 115

5.2.3 Study Procedure 120

5.2.4 Measurements 121

5.3 Study Results and Discussions 124

5.3.1 Demographic Information 124

5.3.2 Effects of robot deception 127

5.3.3 Effects of a robot's embodiment 131

5.3.4 Ethical Implications of other-oriented robot deception 133

5.3.5 Summary 135

5.4 Extended HRI study 136

5.5 Summary 143

VI ROBOT DECEPTION AND ETHICAL ISSUES 147

6.1 Robot Ethics 148

6.2 Ethical implications from the online survey's results 151

6.3 Summary 162

VII CONCLUSION **165**

APPENDIX A — EXPERIMENTAL SETTINGS OF SQUIRREL ROBOT DECEPTION **172**

APPENDIX B — HRI STUDY SUPPLEMENTS **176**

APPENDIX C — ROBOT ETHICS SURVEYS **190**

APPENDIX D — DECEPTIVE ACTION SELECTION IMPLEMENTATION **196**

REFERENCES **202**

LIST OF TABLES

1 Taxonomies of Deception in the Fields of Philosophy, Psychology, and Economics . . . 25

2 Taxonomies of Deception in the Fields of Military, Cyberspace, and Biology . . . 27

3 Three Dimensions for Robot Deception Taxonomy . . . 29

4 Robot Deception Taxonomy with Examples. . . . 31

5 Robot Experiment Results with Convergence Rate . . . 42

6 Deception Generation Types . . . 48

7 General Gesture Primitives (notation: ggp_i) with necessary parameters and body parts in a humanoid robot; ggp_2 is the deictic gesture and all other gesture primitives are iconic gestures. . . . 51

8 Emotional Gesture Primitives (notation: egp_i) with necessary parameters and body parts in a humanoid robot; These emotional gesture primitives are the metaphoric gestures. . . . 51

9 Humanoid Robot's Proxemic Spatial Regions . . . 60

10 Data structure for the adaptation rule . . . 80

11 Exemplar Scenario: Data structure of state S . . . 91

12 Exemplar Scenario: Initial Casebase . . . 95

13 $F_{emotion}$ from basic emotion to $[valence, arousal]$ w/ EARL classification 96

14 Examples of similarity lookup table (Table 15) calculation for the emotion feature . . . 96

15 Similarity score lookup table $M_{emotion}$ for emotional feature . . . 97

16 Exemplar Scenario: Calculating similarity scores and sorting . . . 99

17 Exemplar Scenario: Rules for adaptation . . . 100

18 Exemplar Scenario: Final casebase with the newly updated case . . . 102

19 2 by 2 mixed-subject design . . . 117

20 Demographic Information from 34 HRI study participants . . . 125

21 Task performance: Average number of correct and incorrect answers from the true and deception condition . . . 128

22 NASA's TLX results: Average ratings from Deception and True groups; Scale: 0 (very low) - 21 (very high) 129

23 Task performance: Average number of correct and incorrect answers from the robot and monitor feedback condition 131

24 Self-reported measures: Impressions of a robot or monitor feedback; Average ratings (standard deviation) and p-value 132

25 Extended HRI study Initial Casebase (predetermined) 140

26 Demographic Information from Five extended HRI study participants 141

27 Extended HRI study results from all five participant: task performance (number of correct answers) and deceptive action generation results (number of deceptive feedback) 142

28 Extended HRI study self-reported measures: Impressions to a deceptive robot, average ratings from the five participants 144

29 NASA's TLX results: Average ratings from Extended HRI study's participants; Scale: 0 (very low) - 21 (very high) 144

30 Demographic Information from a web-based Survey Study 154

31 Survey Results Overview (1-strongly disagree, 5-strongly agree) . . . 155

32 Survey Results by Age Groups (1-strongly disagree, 5-strongly agree) 158

33 Survey Results by Technical Levels (1-strongly disagree, 5-strongly agree) 160

34 Survey Results by Prior Robot Experience (1-strongly disagree, 5-strongly agree) 161

LIST OF FIGURES

1 Examples of Animal Deception. 7

2 Fake bus stop in front of a German nursing home [72] 11

3 Situational Conditions for Other-oriented Deception with Examples . 13

4 Examples of Robot Deception. 20

5 Black Eastern Gray Squirrel moving peanuts 33

6 Squirrel's Cache Protection Strategy 34

7 Abstract level of Finite State Acceptors for squirrel robot's behaviors 35

8 Finite State Acceptors for squirrel robot's patrolling strategy in MissionLab 36

9 Robot Experiment Layout 39

10 Robot Platform: Pioneer robot with omni-directional camera 39

11 Robot Experiment Scenario 40

12 Criminology-inspired computational architecture for a robot's other-oriented deception 45

13 Overview of the action generation mechanism via deception transformation layers for nonverbal action cues 48

14 Overview of the action generation mechanism via gesture transformation layer 52

15 Examples of gesture transformation via Deception by Commission mechanism; Gesture primitive pairs which are in the set of gesture primitives pairs P. Therefore, when one of gestures is selected as a true set, the alternate gesture is used as a deceptive gesture according to Equation 2. 55

16 Simulations of deceptive "pointing" gesture generation via Deception by Commission mechanism. According to Equation 3, the alternative object's location is selected as deceptive pointing action position. . . 56

17 Overview of the action generation mechanism via facial expression transformation layer 57

18 Example of deceptive facial cue generation with the R25 robot [126] . 58

19 Overview of the action generation mechanism via proximity transformation layer 59

20 Example for Type 2 (commission) deceptive proximity generation . . 61

21 Detailed integration step of the action generation mechanism (extended from Figure 13) 61

22 Overview of Action Selection Model using CBR 74

23 Overview of final action selection in case adaptation 78

24 Computational architecture for the motives/opportunities model . . . 90

25 The original flowchart of START Triage process; copyright and permission from http://www.remm.nlm.gov/, originally adapted from http://www.start-triage.com/ 94

26 Similarity Score Calculation strategy 98

27 Exemplar Scenario: Case Adaptation Process 101

28 Exemplar Scenario: Casebase Updating Strategy 102

29 Six different pills and two-row pill organizer in Medication sorting task 109

30 Sorting Task Instruction shown on iPad 110

31 3-back auditory task example 110

32 Nao robot platform and its feedback of the participant's performance 111

33 Compensation Guideline 113

34 Experimental Settings 113

35 Sorting Task Instructions Examples 114

36 Robot assistant's feedback of the participant's performance: Happy/Yes gestures indicate the participant's correct answer and Sad/No gestures mean the participant's incorrect answer 116

37 Between-subject conditions: (Left) Feedback without deception, (Right) Feedback with deception 118

38 Non-robotic visual feedback of the participant's performance: A green screen indicates the participant's correct answer and a red screen means the participant's incorrect answer 118

39 Self-reported measure: NASA's TLX is collected after each task session. 123

40 Demographic Information Chart: Technology level, Robot interaction experience, and Education level 126

41 NASA's TLX results from Robot feedback condition: Red-average ratings from Deception group, Green-average ratings from True group . 129

42 Post-survey results: Impressions of robot (red) or monitor (blue) feedback during the task 133

43 Post-survey results; Ethical Question: I can accept robot deception in [Context in x-axis] if it is strictly used only to benefit humans; Scales in y-axis: 1-strongly disagree, 5-strongly agree 134

44 Post-survey results; Ethical Question: I can accept robot deception in [Context in x-axis] if it is strictly used only to benefit humans; Scales in y-axis: 1-strongly disagree, 5-strongly agree (recall from Figure 43); Average ratings to agree to other-oriented robot deception in specific context increase. 152

45 Survey Results Overview: I can accept robot deception in [Context in x-axis] if it is strictly used only to benefit humans; Scales in y-axis: 1-strongly disagree, 5-strongly agree 156

46 Survey Results by Age Groups: I can accept robot deception in [Context in x-axis] if it is strictly used only to benefit humans; Scales in y-axis: 1-strongly disagree, 5-strongly agree 159

47 Survey Results by Technical Level: I can accept robot deception in [Context in x-axis] if it is strictly used only to benefit humans; Scales in y-axis: 1-strongly disagree, 5-strongly agree 161

48 Survey Results by Prior Robot Experiences: I can accept robot deception in [Context in x-axis] if it is strictly used only to benefit humans; Scales in y-axis: 1-strongly disagree, 5-strongly agree 162

49 Finite State Acceptors for competitor robot's hunting strategy in MissionLab 174

SUMMARY

Deception is an essential social behavior for humans, and we can observe human deceptive behaviors in a variety of contexts including sports, culture, education, war, and everyday life. Deception is also used for the purpose of survival in animals and even in plants. From these findings, it is obvious that deception is a general and essential behavior for any species, which raises an interesting research question: can deception be an essential characteristic for robots, especially social robots? Based on this curiosity, this dissertation aimed to develop a robot's deception capabilities, especially in human-robot interaction (HRI) situations. Specifically, the goal of this dissertation is to develop a social robot's deceptive behaviors that can produce benefits for the deceived humans (other-oriented robot deception). To achieve other-oriented robot deception, several scientific contributions were accomplished in this dissertation. A novel taxonomy of robot deception was defined, and a general computational model for a robot's deceptive behaviors was developed based on criminological law. Appropriate HRI contexts in which a robot's other-oriented deception can generate benefits were explored, and a methodology for evaluating a robot's other-oriented deception in appropriate HRI contexts was designed, and studies were conducted with human subjects. Finally, the ethical implications of other-oriented robot deception were also explored and thoughtfully discussed.

CHAPTER I

INTRODUCTION

As social agents, people commonly lie to others and perform deceptive behaviors more than they realize [85]. In human interaction, deception is ubiquitous and occurs frequently during people's development and in personal relationships [16], sports [96], culture [85], and even war [69]. Then, are humans the only beings to have deception capabilities? No, deception is not limited to human beings. Various biological research findings illustrate that animals act deceptively in several ways to enhance their chance of survival [124]. One article finds that even some plants show deception for the purpose of survival [133]. From these findings, we can argue that deception is a general and essential behavior for any species, which raises an interesting question: can deception be an essential characteristic for robots, especially social robots?

Studies on deception in psychology provide some clues for this question. Vasek stated that "the development of deception follows the development of other skills used in social understanding. [167]" Therefore, deception is an essential factor for humans, not just for survival but for humans to be social creatures. Dennett's argument about intentionality also illustrates the important role of human deception. Intentionality is defined as "the power of minds to be about, to represent, or to stand for, things, properties and states of affairs. [44]" Intentionality is thus one of the key factors defining humans as social agents. According to Dennett, a higher-order intentionality can be achieved by adding several different features, with capability for deception notably among them [44]. In summation, we can say that deception capabilities can be an important factor in social intelligence and agency.

The use of social robots is exploding into multiple applications in our everyday

life. For example, companion robots are broadly used in elder care or childcare [125, 61, 81, 173]. Robot assistants have also been introduced in the context of education to increase students' learning efficiency [158, 157]. Recently, social robots such as Pepper [47] have even been promoted to interact with people at home to enhance their lives. By increasing the use of robots in human-robot interaction situations, robots will more frequently play a role as social agents, and naturally, the aim to develop more socially intelligent robots is growing. And as stated above, to achieve more sophisticated social robots, deception should be considered as one of the important factors in robotics research.

Despite the need for robot deception, little research on robot deception has been conducted until now. Especially, in regard to robotic deception, research has focused on specific situations such as military robots' deception [38, 139]. However, a military context involves vulnerable situations, which should be handled differently from our usual social contexts. Throughout the work in this dissertation, it is expected to investigate and explore deception in social robots. In other words, this dissertation aims to figure out whether and how a robot decides and performs deception in general social situations.

Even though we can discuss the potential benefits of robot deception, it is obvious that robot deception has to be considered carefully in regards to social robots. One strong argument that will be illustrated throughout this research is that robot deception should be used only in appropriate human-robot interaction (HRI) contexts. The motivation of robot deception will be discussed later, but briefly, a chief motive for social robots to perform deception should be to benefit the deceived human beings. According to DePaulo [45], human deception can be categorized based on motivation, such as self-oriented and other-oriented deception. In general, people act deceptively for their (deceiver's) own benefit. This is self-oriented deception. However, people also sometimes deceive another person for that person's (the deceived's) benefit. For

example, people may tell a white lie such as "you look great today!" just to make the deceived person feel good. This type of deception is defined as other-oriented deception, which is motivated by the deceived person's potential benefits. More detailed explanations of this categorization will follow in chapters 2 and 3. Inspired by this definition, robot deception will be classified (also the taxonomy of robot deception will be defined) and my own definition of other-oriented robot deception will be provided. Finally, throughout this dissertation, it is aimed to achieve this other-oriented robot deception in HRI.

In sum, the main argument in this research is that social robots' deception should be limited to other-oriented robot deception. In other words, robots' deception can be used only when appropriate HRI contexts contain benefits for the deceived humans. In this dissertation, it will be discussed how a robot's deception can truly produce benefits for humans in social situations, and some models for a robot's other-oriented deception will be also provided. Finally, from the results of this research, I aim to show that a robot, in order to be socially intelligent and interactive, should have deceptive capabilities that will benefit its deceived human partners.

1.1 Research Question

1.1.1 Primary Research Question

The main research question that the work in this dissertation supports is this: ***Can a robot use deception in appropriate HRI domains in order to benefit the deceived human partner?***.

As emphasized in the previous section, this dissertation aims to develop a robot's deception capabilities especially in human-robot interaction situations. In addition, it is strongly argued that a social robot's deceptive behaviors should be only be used when it can produce benefits for the deceived humans. To prove this primary research hypothesis and also develop a model for this benevolent robot deception model, my

research is broken down into five subsidiary questions.

1.1.2 Subsidiary Questions

1. What kinds of deception can be beneficial for those being deceived?

Humans and animals can use different kinds of deceptive behaviors. Similar to humans and animals, robots can also perform deceptive behaviors for specific purposes. Among different contexts, a number of potential situations is investigated in which, deceptive robot behaviors can be beneficial, especially for people being deceived. The aim of this research is to determine the appropriate use of robots' deceptive behaviors to benefit the deceived. Therefore, it is essential to understand in which particular contexts robot deception may benefit deceived people.

2. How can deceptive behaviors be applied to a robotic system?

Deceptive behaviors have to be applied appropriately to robotic systems. Since robots differ from animals and humans in their embodiment and motion/perception capabilities, it will be necessary to determine the most applicable methodologies for robot systems under these limitations and conditions.

3. What formal theoretical/mathematical expressions are appropriate for generating robot deception?

Algorithms should be developed to apply deceptive behaviors to the robot system. Formal theoretical expressions and suitable computational models require development. In particular, the work in this dissertation focuses on the deceptive capabilities of robots in "HRI contexts." Therefore, the development of formal deceptive expressions for a robot while interacting with people is necessary for this research.

4. What are the most effective evaluation methods and metrics to test the research

hypothesis?

After the computational models for generating robot deception are determined and applied to robots, the algorithm must be tested to evaluate if it is truly working. Furthermore, this dissertation aims to address the research hypothesis, which is that robots' deceptive behaviors can benefit deceived people in certain HRI contexts. The hypothesis must be tested to determine whether it is correct according to the specific developed deceptive behaviors for the robot. To answer these questions, it is required to conduct well-designed HRI studies with human subjects as evaluation methods.

5. *How should the ethical issues of robot deception be handled in HRI?*

Even though robot deception can provide several advantages to humans, it is arguable whether deceiving humans is morally acceptable in HRI. Therefore, this ethical issues are also considered thoughtfully in this research.

1.2 Objectives

The main objective of this research is to prove the benefits of robot deception in an appropriate human-robot interaction situation. By answering the primary and subsidiary research questions, I aim to many scientific contributions are accomplished as follows:

- A novel taxonomy of robot deception is defined based on significant literature reviews on deception in a variety of fields, such as psychology, biology, military, economics, and so on.
- A general computational model for a robot's deceptive behaviors is developed based on criminological law.
  - An algorithm to generate a robot's deceptive action is developed and implemented.

  - An algorithm to select an appropriate deceptive action in specific situations is developed and implemented.

- Appropriate HRI contexts in which a robot's other-oriented deception can generate benefits are explored and determined.

- A methodology for evaluating a robot's other-oriented deception in appropriate HRI contexts is designed, and studies are conducted with human subjects.

- The ethical implications of robot deception are explored and thoughtfully discussed.

1.3 Dissertation Outline

Chapter 1 has outlined the main goal of this research, with its primary and subsidiary research questions and contributions. In chapter 2, previous research on deception in various fields, including psychology, biology, and robotics, is reviewed. By using this information from other fields, a novel taxonomy of robot deception is created and introduced in chapter 3. A general computational frameworks of a robot's other-oriented deception is then introduced in chapter 4, and evaluation methods and experimental results follow in chapter 5. It is also essential to discuss ethical issues in robot deception research. Therefore, chapter 6 introduces the ethical implications of robot deception based on survey results. Finally, chapter 7 presents concluding remarks.

CHAPTER II

RELATED WORK

In this chapter, literature related to deception in a variety of fields will be reviewed. For a better understanding of deception, deception in animals and humans is first reviewed, highlighting deception in biology (section 2.1) and psychology (section 2.2). Previous work in robot deception follows in section 2.3. Using deceptive behaviors obviously leads to ethical arguments, even in human cases. Therefore, a discussion of ethical issues in robot deception is an essential part of this research. For this consideration, moral theories and ethical approaches to deception are also reviewed in section 2.4.

2.1 Animal Deception

Animals use various forms of misinformation. These deception mechanisms, achieved by sending false signals either intentionally or unintentionally, are essential for the

(a) Teratodus Monticollis grasshopper is mimicking a leaf.[1]

(b) Killdeer is showing broken wing broken wing act.[2]

(c) Chimpanzees produce complex and intentional deceptive behaviors.[3]

Figure 1: Examples of Animal Deception.

[1] http://en.wikipedia.org/wiki/Camouflage

[2] http://en.wikipedia.org/wiki/Distraction_display

[3] http://en.wikipedia.org/wiki/Chimpanzee

animals' survival. For example, camouflage and mimicry are used by many species (Figure 1(a)). By resembling other animal species or inanimate objects, animals transmit misinformation to others so that they can avoid detection by both predators and their prey. While camouflage or mimicry are examples of unknowingly deceiving, a deceptive behavior can include seemingly more intentional misinformation.

More intentional deceptive behaviors are observed from different animals ranging from insects to primates. The spider genus Portia, which preys primarily on other spiders, deceives its prey by vibrating the web in ways that resemble a small insect getting ensnared. When the resident spider of the web comes to investigate the insects, Portia preys on it [183].

According to Ristau's research [124], another interesting deceptive behavior appears in piping plovers. These birds exhibit a "broken-wing display" deceptive behavior. By feigning an injured wing and hopping farther and farther from the nest, birds lead the predator away from their young, thus protecting them (Figure 1(b)).

Feigning death is another well-know form of animal deception. Mainly, animals appear being dead to defend from predators since many animals only take prey that are living [102, 113]. Besides the purpose of protection from predators, some animals also use this form of deception to improve reproduction and predation. For example, in the spider species Pisaura mirabilis, female spiders generally eat the male during mating, and to avoid getting eaten, male spiders feign death [66]. The predatory Nimbochromis fish lies down on the bottom sediments to appear as a dead fish and then attract scavengers fishes [98].

Primates are the species most commonly ascribed with the ability to deceive [31, 62]. For example, chimpanzees have multiple deceptive behaviors with several different objectives (Figure 1(c)). When chimpanzees find fruit, they do not move directly so that they do not give any indication to competitors that they have noticed the location of the food. Deceptive behavior of chimpanzees is also observed during

interactions with humans. According to one observation, a chimpanzee feigned having his arm stuck in the bars of his cage in order to lure a zookeeper nearby. As soon as the human entered to help free his arm, he leapt onto the zookeeper [40]. More sophisticated examples have been also observed in great apes. A female gorilla has been trained to use America Sign Language and at the end she could use the signs to express her intentions. One day, she tore off a steel sink and surprisingly she started to lie to her handlers by signing "cat did it" and pointing at the cat [63].

Another relevant class of deceptive behavior occurs in the food-hoarding strategies of animals. Food hoarding (caching) is an important type of animal behavior needed for their survival through periods when nourishment is not readily available. In particular, these caching behaviors are commonly observed in rodents, such as hamsters or squirrels [75]. After hoarding the food, animals also patrol the caching locations to protect their food. According to the biological findings, interesting deceptive behaviors are also observed. In this patrolling strategy to protect their food caches from other predators. Eastern Grey Squirrels' behavior is one interesting example in nature regarding the possible role of deception [155]. During the patrolling phase, the squirrel spends time visiting stocked food caches. It was observed, however, that when a predator was present, the squirrel changed its patrolling behavior to spend time visiting empty (fake) cache sites, with the apparent intent to mislead the raider into the belief that those sources were where the valuables were located, a diversionary tactic of sorts.

2.2 Human Deception

Human deception requires extensive planning and second-guessing compared to the planning and deception that most animals are capable of. Many psychologists have discussed human deception from various perspectives. According to Vasek [167], the development of deception follows the development of other skills used in social

understanding such as perspective-taking, communicational/linguistic skills, and understanding of one's own and other's intentionality. In other words, deception is one of the good indicators of the human's Theory of Mind mechanism. From this perspective, the capabilities of the deceptive behaviors have been also discussed for use to determine children's developmental disabilities such as autism [16].

One type of interesting human deception happens in sports. Work by Mawby and Mitchell [96] showed several principles used in sports to enact or avoid deception. For example, many players fake out opponents, thereby redirecting the opponents' actions, and teams also use complicated tactics of deception that require the coordinated actions of several players. Another recent work has analyzed the anticipation skills of deceptive movement in sports [73, 150, 56]. Based on the analysis results, they attempted to predict an opponent's correct direction quickly and exactly. Finally, these kinds of research enable players to train using virtual agents, which have the capability to anticipate. As a result, players can improve their deception skills in rugby, soccer, and handball [21, 46, 170].

Sun Tzu stated in The Art of War [156], "All warfare is based on deception." Actually, people have used deception in warfare to cloak their intentions and movements [69, 99, 58]. In the military sense, the term "deception" is applicable to "any planned measure undertaken for purposes of misleading or deceiving the enemy [134]." Different from other human deception, a deception "story" is an essential instrument for executing a military deception. Here, "story" means a detailed scenario of "that which you want the enemy to think in order to make him do what your commander wants him to do [134]." Because the military utilizes relatively fixed scenarios with histories of thinking and acting, military deception is feasible and frequently implemented.

In animal cases, deception is defined as "a false communication that tends to benefit the communicator [20]." In other words, animals usually act deceptively for their

Figure 2: Fake bus stop in front of a German nursing home [72]

own benefits. An interesting aspect of human deception is that people also act deceptively for the benefit of the deceived party [85, 15, 9]. Inspired by those differences, DePaulo defined a taxonomy of human deception [45]. According to his research, human deception can be mapped into two categorizes, which are self-oriented deception and other-oriented deception. People generally perform deceptive behaviors for their own benefits, and this kind of deception is defined as ***self-oriented deception***. However, people also sometimes show deceptive behaviors for the deceived other's advantages, this is classified as ***other-oriented deception***. DePaulo's study showed that people generally perform more self-oriented deception than other-oriented deception in their everyday lives. However, people still generated other-oriented deception. For example, the participant said his friend's cookies are delicious even though he didn't think so to protect his friend from feeling bad, or simply to make his friend feel good. This kind of other-oriented lies/deception happen in people's everyday lives [45].

Other-oriented deception happens frequently for medical purposes. For example, the use of placebos is to benefit patients who are deceived by doctors/nurses in medicine [100]. In front of a German nursing home, a fake bus stop is located to deceive Alzheimer's patients [72]. These patients sometimes wander off and go to the bus stop to go back home. By having this fake bus stop, those Alzheimer's patients

can be protected from risky situations (Figure 2). Sometimes, deception can benefit patients by reducing fear and improving the results of healthcare. In one case, a caregiver tried to deceive developmentally disabled adults who were afraid of bleeding gums by giving them red-colored toothpaste. With this deception, the patients were able to improve their dental condition since they would brush more than they did before.

In a crisis, victims' emotional state can seriously affect their safety [179]. Also, when victims' cooperation is required during Search and Rescue tasks, managing their emotions is important. For this reason, rescuers sometimes hide the truth of the situation and act deceptively such as not describing the severity of injuries or the situation to victims accurately [89].

We can also observe some other-oriented deception-related concepts in the education domain. One interesting theory is the Pygmalion effect [130]. According to Robert Rosenthal and Lenore Jacobson's study, students' performance and learning efficiency can be increased when teachers have higher expectations for the students. Therefore, when teachers deceptively show greater expectations the students may be motivated to increase their learning efficiency.

More generally, we can also observe many other-oriented deceptions for different purposes in our everyday lives. Many people tell white lies for the deceived individual's feelings or benefits. For example, people sometimes lie to a friend that "you look so good" just to make the friend feel better. A surprise party can also be considered an other-oriented deception since people hide the party information to maximize the deceived person's happiness. Magic or jokes also sometimes use deception techniques for entertainment purposes.

Observing other-oriented deception in psychology is essential prerequisite work in my research since I hypothesize that robots' deceptive capabilities can benefit the deceived human partners similar to these same capabilities in human cases. By deeply

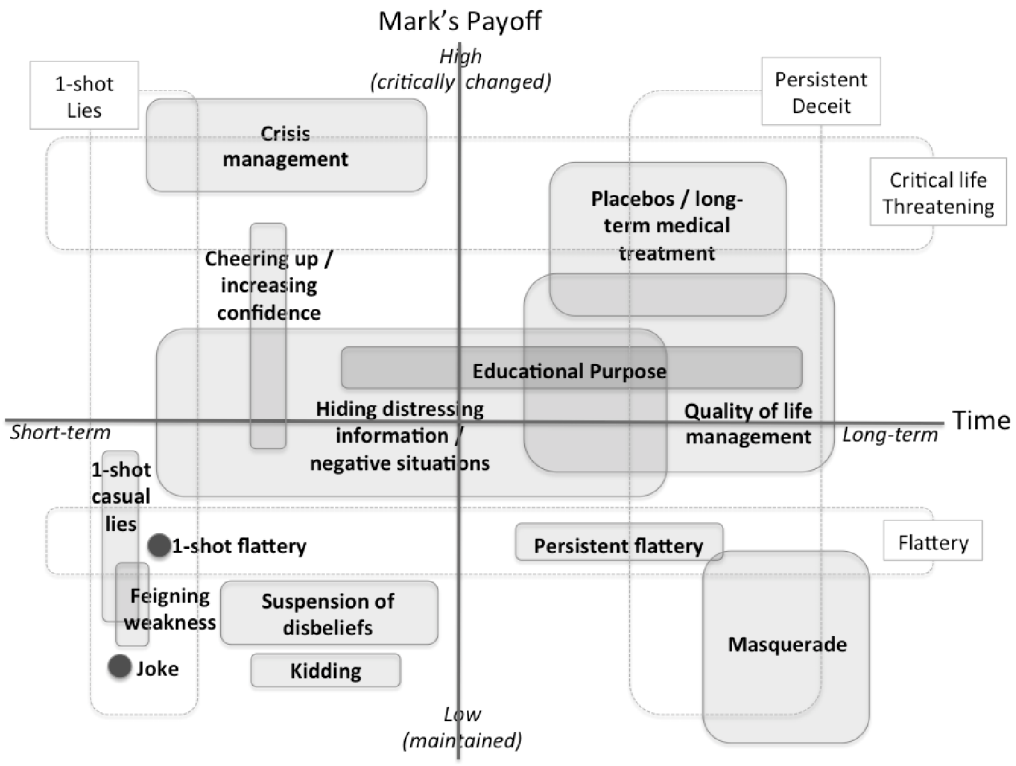


Figure 3: Situational Conditions for Other-oriented Deception with Examples

understanding the features and uses of other-oriented deception in human-human interaction, I believe we can also find some clues and basis for the use of other-oriented deception in human-robot interaction contexts. To understand humans' other-oriented deception more specifically, I grouped relevant situations and defined the situational conditions pertaining to the application and utility of their other-oriented deception. From reviewing various situations when other-oriented deception occurs between humans, I was able to group these situations along two dimensions: 1) the time duration of the deception and 2) the payoff of the deceived person (the mark). The time dimension ranges from one-shot to short-term to long-term, referring to the length of time deception is maintained by the deceiver's actions. The deceived person's payoff is categorized by the effect on the deceived person's outcome (ranging from high to low payoff).

As shown in Figure 3, representative other-oriented examples in these dimensions

are illustrated by their location in this two dimensional space. Situational conditions include the following examples.

1. *Crisis management* is a situation where the deceiver's deceptive behaviors or lies must have a rapid effect on the mark (short-term) perhaps in a life-threatening high payoff situation. For example, other-oriented deception in a search-and-rescue situation may involve immediate emotional or physiological remediation for a victim [89]. Lying to a mark regarding the direness of their situation in order to calm him/her down or to increase their confidence may increase their likelihood of survival in this life critical situation.

2. When someone faces a highly stressful situation such as a big presentation in front of huge crowd or an athletic trial, people sometimes lie to *cheer up / increase their confidence* to let the speaker calm down in the short-term such as "Don't worry! You're perfectly prepared" or "I know you can successfully do this."

3. *Quality of Life Management (QoLM)* involves maintaining deception over long periods of time, again for potential life-critical (health) situations in therapeutic treatment of serious or generative illness, or regarding status of long-term economic well-being. For example, *placebos* may be persistently used for a deceived patient's long-term benefit [100]. Long-term lying can also be used in a similar manner with the hopes of benefitting the patient.

4. Sometimes, teachers also behave deceptively or lie for *educational purposes*, perhaps playing dumb for example [130, 94]. This deception can increase the student's learning efficiency, and it produces long-term benefits to the deceived person, although the deceit may be either short or long-term.

5. One-Shot *Casual Lies* are common in general conversation [45]. Generally, deceivers act deceptively or tell a lie to maintain the deceived other's emotional state for good. For example, general lies such as "you look nice today" or "I like your clothes" are obvious examples of 1-shot casual lies. These are not life-critical

situations. "That was a great presentation" can also be another such example.

6. *Flattery* also ranges from the short-term to long-term to make the deceived other's emotional state beneficial to their performance. *Persistent flattery* is an example (e.g., "a**-kissing") that makes the deceived person feel undeservedly better about themselves for a relatively long-term period. In this long-term case, benefit (payoff) accrues for both the deceived person and the deceiver, but I focus for now solely on the benefits to the mark.

7. People sometimes *feign weakness* to make marks feel better by helping deceivers in short-term periods. The deceived person can maintain emotionally good state or feel better and confident from this deception. For example, a woman might pretend not to be able to open a jar just to make the man feel better and more confident about himself.

8. One-shot *Jokes* or more persistent *kidding* using deception is also an example of short-term lies, since they aim to maintain a good atmosphere of social community by making the deceived person feel at ease perhaps by stating falsehoods about themselves, others, or a situation in a humorous and non-truthful way.

9. Promotion of *suspension of disbelief* uses deception to provide the deceived person with fun and enjoyment. For example, movies, magic, or other fictional works use illusion to deceive others. This differs from other examples of deception, since the deceived others voluntarily allow themselves to be deceived.

10. A *masquerade* is characterized by deception that persists for extended periods of time to create an illusion regarding something that does not exist, but may make the mark feel better about themselves.

11. Sometimes, people *hide distressing information or negative situations* from others, assuming they may be able to resolve it on their own without additional help from the deceived person and so not induce anxiety in the deceived.

In sum, we can observe many human cases of other-oriented deception as described in this section, and the research in this dissertation also aims to develop a robot's other-oriented deception to benefit the deceived humans in an appropriate HRI context similar to those other-oriented deception in human cases.

2.3 Robot Deception

Endowing robots with the capacity for deception has significant potential utility [176], similar to its use in humans and animals. As stated in 2.2, deceptive behaviors are useful in the military domain [69, 99]. Similar to human cases, military robots' capable of deception could mislead opponents in a variety of ways. As both individual and teams of robots become prevalent in the military's future [164, 134], robotic deception can provide new advantages apart from the traditional one of force multiplication. In other areas, such as search and rescue or healthcare, deceptive robots might also add value similar to human cases, for example, for calming victims or patients when required for their own protection. Conceivably, even in the field of educational robots, the deceptive behavior of a robot teacher may potentially play a role in improving human learning efficiency. Despite the ubiquity in nature and the potential benefits of deception, very few studies have been done on robot deception to date.

One interesting application in robot deception is the camouflage robot, developed at Harvard University [103]. Camouflage is a widely used deception mechanism in animals and militaries. Inspired by these real-world usages, the researchers at Harvard developed this soft robot, which can automatically change the color of body to match its environment.

Motion camouflage has also been studied for robot systems. Unlike the previous type of camouflage, motion camouflage is a behavioral deception capability observed in dragonflies. By following indirect trajectories, dragonflies can deceptively approach as if they were remaining stationary from the perspective of the prey. Carey et

al. developed an optimal control mechanism to generate these motion camouflage trajectories and verified it with simulation results [27]. For real robot systems, more recent research proposed new motion camouflage techniques that are applicable to unicycle robots [120].

Floreano's research group [55] demonstrated robots evolving deceptive strategies in an evolutionary manner, learning to protect energy sources. Their work illustrated the ties between biology, evolution, and signal communication and does so on a robotic platform. They showed that cooperative communication evolves when robot colonies consist of genetically similar individuals. In contrast, when the robot colonies were dissimilar, some of the robots evolved deceptive communication signals.

Wagner and Arkin [176] used interdependence theory and game theory to develop algorithms that allow a robot to determine both when and how it should deceive others. Recent work at Georgia Tech explored the role of deception according to Grafen's dishonesty model [77] in the context of bird mobbing behavior [38]. Another study developed robots' deceptive behavior inspired by squirrel's deceptive food protection behaviors and showed how a robot successfully uses this deception algorithm for resource protection [139].

By increasing the use of robots in human life, the development of social robots is also getting more important. Many techniques for better HRI has been developed such as emotional intelligence [114, 23, 11], collaborating between robots and humans [135, 132, 53], social learning [8, 32], turn-taking and engaging mechanisms [107, 29], assistive technologies [110], and so on. In addition, many researchers aim to build more socially-intelligent robots that feature intentionality. Here, intentionality means "the power of minds to be about, to represent, or to stand for, things, properties and states of affairs. [44]" Therefore, by having intentionality, robots can interact with human partners more naturally and effectively. According to Dennett [44], a high-order intentionality can be achieved by adding several different features, notably

deception capability. In addition, deception is highly related to the theory of mind model since it requires to anticipate and manipulate other's actions [19]. In sum, we can argue that more intentional and autonomous social robots are possible when deception capabilities are added.

Based on this argument, much research on robot deception has also been proposed in HRI contexts. Terada and Ito [160] demonstrated that a robot is able to deceive a human by producing a deceptive behavior contrary to the human subject's prediction. These results illustrated that an unexpected change of the robot's behavior gave rise to an impression in the human of being deceived by the robot.

Other research shows that robots' deceptive behavior can increase users' engagement in robotic game domains. Work at Yale University [146] illustrated increased engagement with a cheating robot in the context of a rock-paper-scissors game. They proved greater attributions of mental state to the robot by the human players when participants played against the cheating robots than the true robots. At Carnegie Mellon University [169] a study showed an increase of users' engagement and enjoyment in a multi-player robotic game in the presence of a deceptive robot referee. By declaring false information to game players about how much players win or lose, they observed whether this behavior affects a human's general motivation and interest based on frequency of winning, duration of playing, and so on. These results indicate that deceptive behaviors are potentially beneficial not only in the military domain but also in a human's everyday context.

Brewer et al. showed that deception can be used in a robotic physical therapy system [24]. By giving deceptive visual feedback on the amount of force patients currently exert, patients can perceive the amount of force lower than the actual amount. As a result, patients can add additional force and get benefits during rehabilitation.

Research from the University of Tsukuba [94] showed that a deceptive robot partner can improve the learning efficiency of children. In this study, the children participated in a learning game with a robot partner, which pretends to learn from children. In other words, the robot partner in this study is a care-receiving robot, which enables children to learn by teaching [158]. The goal of this learning game is for kids to draw the shape of corresponding English words such as circle, square, and so on. The interesting part is that the robot acted as an instructor, but deliberately made mistakes and behaved as if it did not know the answer. According to the results, by showing these unknowing/unsure behaviors, the learning efficiency of the children was significantly increased. Since robots' unsure/dumb behaviors can affect a human's learning efficiency, I assume that these results relate to a robot's deceptive capabilities. As a result, I can conclude that this study provides preliminary results of the positive effects of robots' deceptive behavior in education contexts.

Westlund and Breazeal introduced an interesting research hypothesis based on their preliminary HRI study in child-robot interaction [178]. While experimenters remotely controlled the robot, they deceptively told the children that the robot was autonomously acting and also controlled to appear as an autonomous agent. The preliminary results led to the hypothesis that children stick to the robot more and disclose their secrets to the robot agents more than to their parents or teachers.

Recently, researchers examined a robot's deceptive goal-directed motion using user studies [48]. By analyzing the results, they presented human strategies/reactions to robot deception, mathematical models for deceptive motion generation, and other implications of robot deception.

Nowadays, robots are broadly used as companions or care partners in elder care or childcare [157, 174]. And sometimes, the elderly or children are deceived into thinking that the robots are social being or real pets. Some researchers argue that this is also a form of deception since the robots pretend to be some other entity [35, 95]. In fact,

(a) KASPAR robot[4]

(b) Probo robot[5]

(c) Pleo robot[6]

(d) Paro robot[7]

Figure 4: Examples of Robot Deception.

several case studies illustrated that children perceived care robots as emotional and social beings like humans. Tanaka introduced a Qrio robot as a peer to children in his long-term study, and the children believed that the robot was a social being [157]. Other social robots such as Kaspar (Figure 4(a)) [125], Probo huggable robot (Figure 4(b)) [61, 147], or Pleo robot (Figure 4(c)) [81] were also introduced to children, including children with special needs, as if they are peers or pets rather than robots. The healthcare robot Paro (Figure 4(d)) is also positively used in the care of dementia patients by introducing it as a real pet [173, 174].

[4] http://www.herts.ac.uk/kaspar/

[5] http://http://probo.vub.ac.be/

[6] https://en.wikipedia.org/wiki/Pleo

[7] https://en.wikipedia.org/wiki/Paro_(robot)

2.4 Ethical Theory

Despite the potential benefits of robot deception, relatively little research has been conducted to date on this topic, perhaps as a result of ethical considerations involving this somewhat controversial topic. Therefore, as stated in the subsidiary research questions, reviewing and arguing ethical issues are essential in this research.

Robot ethics is a rapidly expanding area [136, 51]. Especially, many ethical questions can arise when deception is applied to the robotic system [87, 177]. For example, we can face ethical questions such as "Is deception acceptable even in humans?" or "Should a robot be allowed to lie?" Furthermore, since deception is relate to trust [65], the discussion of deception is getting more important.

To discuss ethical issues of robot deception, it is necessary to review the fundamental moral theories of deception in philosophy. According to Kantian theory, deception or lies should always be prohibited, a standard outcome of any ethics classroom in the application of the Categorical Imperative [34]. By this standard, any deceptive behaviors or lies are morally incorrect, human or robot. The utilitarian perspective, on the other hand, argues that an action is morally right and acceptable if it leads to increasing total happiness over all relevant stakeholders [149]. By this perspective, it can also be argued that if deception increases the total benefits among the involved relationships, it is ethically correct [149, 34]. More specifically, Bentham and Mill [148] argued that it is morally right if and only if any behaviors/acts produce overall increased happiness. In other words, an action is morally good if it provides overall benefits. This ethical theory is called act-utilitarianism.

Related to robot deception, Reynolds and Ishikawa [122] discussed the role of designers and robots and emphasized the importance of morally responsible entities. Arkin [13] also pointed how important it is in discussing the ethical justification of robot deception.

In affective computing, researchers have argued that the use of emotional robots

is deceptive. According to Coeckelbergh, emotional robots are deceptive since "1. Emotional robots intend to deceive with their "emotions." 2. Robotic emotions are unreal. 3. emotional robots pretend to be a kind of entity they are not. [35]" For example, when such robots are used in elder care, peoples are led to believe that they are loved or cared for by the robots, and according to the definition, this can be a case of delusion [154]. Finally, by reformulating these three claims of emotional robot deception to ethical criteria, he argued that the situation can also be considered "ideal emotional communication" rather than deception.

More recently, Mattihias suggested four criteria for robot deception [95]. He argued that by fulfilling four conditions, robot deception can be morally permissible. These four criteria are trust, autonomy, transparency, and safety. First, robot deception should not betray patients' trust by promoting patients' interests. Also, deception should support patients' autonomy by supporting them to make decisions and control the machines better. To be transparent to the patients, the fact that deception is happening should be suggested at some point in the conversational context. Finally and most importantly, deception should not lead to any harm to the patients.

There is also an HRI study related to the human moral stance and robot deception. In Kahn's HRI study [79], subjects were asked to play a game, and a humanoid robot guided and observed their performance. After completing the game, a robot debriefed the subjects but announced their achievement deceptively as being lower, and here, researchers observed that many people held the robot morally accountable.

Researchers also sometimes argue ethics of robot deception based on situations. Nijholt identified the potential situations or contexts in which artificial human partners will be deceptive (or not honest) by analyzing human-human cases [108]. He proposed four categories of situations in which natural deceptive interactions will be used or required in human-robot or human-computer interactions. These four categories are 1) conversations and dialogues; 2) commerce, negotiation, persuasion; 3)

teaching, training, serious games; and 4) sports, games, and entertainment. There are also several ethical arguments about robot deception in different contexts, including economics and law [68], healthcare [95], and so on.

The main purpose of this discretion is not to resolve this ethical argument of robot deception entirely. Practically, in robotics, it is even more complicated to state the ethical issues related to deception than in human cases. However, it is obvious that this issue should be carefully and thoughtfully considered while robot deception is developed and applied [10], and therefore, it will be an integral part of my research and discussed further in chapter 6.

CHAPTER III

A TAXONOMY OF ROBOT DECEPTION

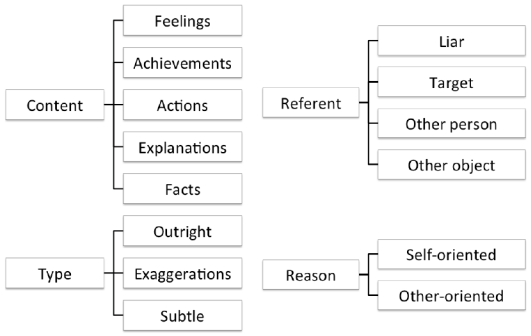
This dissertation argues that the use of deceptive capabilities in robotics features many potential benefits. As shown in the literature review, some robotics research has been proposed and developed related to robot deception, and this topic is becoming an important and interesting research question. However, much of the current research on robot deception focuses on applications and not on fundamental theory such as the delineation of a taxonomy for robot deception. I contend that defining robot deception and establishing a taxonomy are important as a foundation for further robotics research on the subject, and herein such a taxonomy is defined. To accomplish this goal, different ways of defining deception from the different fields are first carefully review (Section 3.1) . Based on the literature reviews, I propose a novel way to define a taxonomy for robot deception (Section 3.2). To show how this taxonomy exactly applies for a specific application, one of my previous works on robot deception is introduced and it is classified according to my taxonomy of robot deception as an example (Section 3.3).

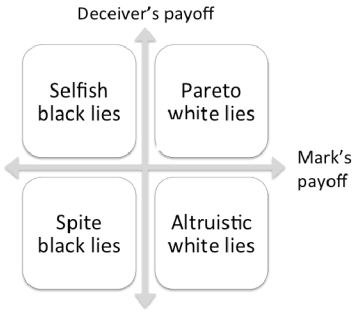
3.1 Taxonomies of Deception from a Human Perspective

In other disciplines, researchers have developed the definitions and taxonomies of deception drawing from the fields of psychology, military, engineering, and so on. In this section, several ways to define and categorize deception in different fields are reviewed followed by a suggested taxonomy of deception from a robotic perspective.

Several definitions of deception have been proposed in different fields. According to Vrij [172], deception is "A successful or unsuccessful deliberate attempt, without forewarning, to create in another a belief that the communicator considers to be

Table 1: Taxonomies of Deception in the Fields of Philosophy, Psychology, and Economics

<table>
<tr><th>Field</th><th>Method</th><th>Taxonomy</th></tr>
<tr><td>(a) Philosophy</td><td>Logical and philosophical view points with a proposition</td><td>
<table>
<tr><th colspan="2"></th><th>Omission (O)</th><th>Commission(C)</th></tr>
<tr><td rowspan="2">Intended (I)</td><td>Pos (P)</td><td>O-P-I</td><td>C-P-I</td></tr>
<tr><td>Neg (N)</td><td>O-N-I</td><td>C-N-I</td></tr>
<tr><td rowspan="2">Unintended (U)</td><td>Pos (P)</td><td>O-P-U</td><td>C-P-U</td></tr>
<tr><td>Neg (N)</td><td>O-N-U</td><td>C-N-U</td></tr>
</table>
</td></tr>
<tr><td>(b) Psychology</td><td>Analysis results of diary studies and surveys</td><td>


</td></tr>
<tr><td>(c) Economics</td><td>Deceiver's and mark's consequences</td><td>


This chart is reproduced with permission from [54]'s Figure 1. Taxonomy of Lies on Change in Payoff.
</td></tr>
</table>

untrue in order to increase the communicators' payoff at the expense of the other side." De Waal stated that "Deception can be defined as the projection, to one's own advantage, of an inaccurate or false image of knowledge, intentions, or motivations" in the paper [40]. We can find a simpler definition of deceptive behavior from a paper by Bond and Robinson [20] who defined it as "a false communication that tends to benefit the communicator."

Taxonomies of deception have been studied extensively by observing different human cases. Several ways to categorize deception have been proposed already by different psychologists and philosophers. Chisholm and Freehan [33] categorized deception from a logical and philosophical viewpoint. Three dimensions were described for distinguishing among types of deception such as commission-omission (the attitude of the deceiver; the deceiver "contributes causally toward" the mark's changes or the deceiver "allows" the mark's changes with respect to belief states), positive-negative (the belief state of the mark; the deceiver makes the mark believe that false proposition is true vs. true proposition is false), and intended-unintended (whether the deceiver changes the mark's belief state or merely sustains it). From the combination of those three dimensions, they provided eight categories of human deception as shown in Table 1(a).

From the results of diary studies and surveys, DePaulo [45] divides deception in four different ways: content, type, referent, and reasons (Table 1(b)). Subcategories of these kinds of deception are also observed and defined. Subcategories of content are feelings, achievements, actions, explanations, and facts. In the category of reasons, there are subcategories of self-oriented and other-oriented deception. Self-oriented deception is used for the deceiver's own advantages. Conversely, other-oriented deception is motivated by the benefits that accrue to the person who is being deceived. In type category, outright, exaggerations, and subtle were defined as subcategories. Also, four different referents were suggested such as liar, target, other person, and

Table 2: Taxonomies of Deception in the Fields of Military, Cyberspace, and Biology

<table>
<tr><th>Field</th><th>Method</th><th>Taxonomy</th></tr>
<tr><td>(a)
Military</td><td>Representing deception</td><td>• Dissimulation
  • Masking: hiding in background
  • Repacking: hiding as something else
  • Dazzling: hiding by confusion
• Simulation
  • Mimicking: deceiving by imitation
  • Inventing: displaying a different reality
  • Decoying: diverting attention</td></tr>
<tr><td>(b)
Cyberspace</td><td>Semantic cases (Linguistic case theory)</td><td><table>
<tr><td>Space</td><td>Direction, location-at, loc-from, loc-to, loc-through, orientation</td></tr>
<tr><td>Time</td><td>Frequency, time-at, time-from, time-to, time-through</td></tr>
<tr><td>Participant</td><td>Agent, beneficiary, experiences, instrument, object, recipient</td></tr>
<tr><td>Causality</td><td>Cause, contradiction, effect, purpose</td></tr>
<tr><td>Quality</td><td>Accompaniment, content, manner, material, measure, order, value</td></tr>
<tr><td>Essence</td><td>Super type, whole</td></tr>
<tr><td>Speech-act</td><td>External precondition, internal precondition</td></tr>
</table></td></tr>
<tr><td>(c)
Biology</td><td>Cognitive complexity</td><td>Intentional vs. Unintentional Deception</td></tr>
</table>

object/event.

Erat and Gneezy [54] classified four types of deception based on their consequences: selfish black lies, spite black lies, pareto white lies, and altruistic white lies, and evaluated it using the human-subjected experiment. For example, as shown in Table 1(c), deception that increases both the deceiver's and mark's (the deceived person's) benefits is classified as "pareto white lies." However, if deception only increases the mark's payoffs but decreases the deceiver's payoffs, it is known as "altruistic white lies." Conversely, "selfish black lies" and "spite black lies" tend to decrease the mark's payoffs while it is distinguished by the deceiver's positive or negative payoffs.

The military is one of the biggest contexts for the use of deceptive behavior. Dunnigan and Nofi [49] proposed a taxonomy of deception based on ways to generate deceptive behaviors. Whaley [180, 18] suggested six categories of deception and grouped them into two sets (Table 2(a)). The six categories of deception are masking, repackaging, dazzling, mimicking, inventing, and decoying. These categories are grouped into dissimulation and simulation. The first three, masking, repacking and dazzling, are categorized as dissimulation (the concealment of truth) and the others are in the simulation category (the exhibition of false).

In cyberspace, deception happens frequently and a taxonomy of deception has been proposed by Rowe et al. [131] for this domain. They defined seven categories of cyberspace deception based on linguistic case theory, including: space, time, participant, causality, quality, essence, and speech-act. By exploring subcategories on each case, they proposed 32 types for a taxonomy of cyberspace deception (Table 2(b)).

Many deceptive behaviors are also observed in nonhuman cases. Animal deception can be categorized depending on its cognitive complexity [43], specifically the two categories of unintentional and intentional animal deception (Table 2(c)). Unintentional animal deception includes mimicry and camouflage. In contrast, intentional deception requires more sophisticated behavioral capacities such as broken-wing displays or in many non-human primate examples such as chimpanzee communication [39].

Recently, researchers in human-computer interaction (HCI) defined the notion of *benevolent deception*, which aims to benefit not only the developers but also the users [5]. They have not proposed a taxonomy of deception, but provided new design principles regarding deception in HCI.

Table 3: Three Dimensions for Robot Deception Taxonomy

Dimensions	Categories	Specifications
Interaction Object	Robot-human deception (H)	Robot deceives human partners
	Robot-nonhuman deception (N)	Robot deceives nonhuman objects such as other robots, animals, and so on.
Interaction Goal	Self-oriented deception (S)	Deception for robot's own benefit
	Other-oriented deception (O)	Deception for the deceived other's benefit
Interaction Type	Physical/ unintentional deception (P)	Deception through robot's embodiments, low cognitive / behavioral complexity
	Behavioral/ intentional deception (B)	Deception through robot's mental representations and behaviors, higher cognitive complexity

3.2 A Taxonomy of Robot Deception

Based on the reviews of the definitions and taxonomies in other disciplines, I developed a taxonomy of robot deception [140]. Similar to human and animal deception, robot deception happens during the interactions among robots or between humans and robots. Therefore, analyzing these "interactions" can identify the key factors to categorize robot deception. Similar to Chisholm and Freehan's approach [33], the salient dimensions of robot deception are first specified, and then the taxonomy of robot deception is defined based on these characteristics.

The three dimensions of robotic deception from the aspect of interactions are interaction object, interaction goal, and interaction method (Table 3).

1. Interaction Object: The interaction object indicates with whom the robot interacts and tries to deceive. In this category, robot deception can be classified into deception in robot-human interactions and in robot-nonhuman interactions.

2. Interaction Goal: This approach is similar to the distinctions of DePaulo's taxonomy, especially his "reason" category [45]. In other words, robot deception

is categorized based on the reason why a robot tries to deceive others: self-oriented deception or other-oriented deception. Self-oriented deception means that a robot's deceptive behaviors benefit the robot itself. In contrast, other-oriented deception happens when the goal of robot deception is to give the advantage to the deceived robots or human partners.

3. Interaction Method: This dimension refers to the way the robot generates deception and it is similar to the taxonomy of animal deception: intentional and unintentional deception. It includes embodiment/physical deception and mental/behavioral deception. Embodiment/physical deception indicates robot deception from morphologies such as camouflage robots. In mental/behavioral deception, a robot generates more intentional deceptive behaviors and represents these behaviors using several cues, which are already included in each robot system.

As a result, I provide three dimensions for distinguishing among the types of deception such as human/nonhuman, self-oriented/other-oriented, physical/behavioral. From the combinations of those three dimensions, a taxonomy of robot deception is defined as shown in Table 4. Each type consists of combination of three characteristics, leading to eight different types of robot deception. As shown in this table, N-S-P and N-O-P types do not have specific examples in robot contexts yet. Therefore, these two types can be excluded in this definition of taxonomies for now. In other words, based on the characteristics of interactions in current robot deception systems, six different usable types of robot deception are defined in the taxonomy. However, there always exist possibilities that those types of robot deception are developed in the future work.

The table also includes examples of each type of robot deception. The camouflage

Table 4: Robot Deception Taxonomy with Examples.

Taxonomy	**Definition**	**Examples**
H-S-P	Deceiving human for deceiver robot's own benefit using physical interactions	Camouflage robots - DARPA's soft robot [103]
N-S-P	Deceiving other robot or nonhuman for deceiver robot's own benefit using physical interactions	N/A
H-O-P	Deceiving human for deceived human's benefit using physical interactions	Android Robots
N-O-P	Deceiving other robot or nonhuman for deceived other's benefit using physical interactions	N/A
H-S-B	Deceiving humans for deceiver robot's own benefit using behavioral interactions	Robot deception in HRI [160]
N-S-B	Deceiving other robots or nonhumans for deceiver robot's self benefit using behavioral interactions	Mobbing robot [38], Robot deception using interdependence theory [176], Squirrel-like robot deception [139]
H-O-B	Deceiving humans for deceived human's benefit using behavioral interactions	Robot deception in entertainment [146], Deceptive robot learner [94], Robot referees in game [169]
N-O-B	Deceiving other robots or nonhumans for deceived other's benefit using behavioral interactions	Robot Sheepdog [168], RoboSquirrel [78]

robot change its color to hide from human's observation. According to my categorization, it can be analyzed that the robot tries to deceive human partners (H) to get its own benefit (S) by transforming its physical appearance (P). Finally, I can classify its category as H-S-P in my taxonomy. Many deceptive robot behaviors that have been developed for the military purposes can be categorized in N-S-B or H-S-B since those robots aim to deceive human or robot opponents by changing behavior patterns (e.g., misleading directions) for producing own benefits [38, 139]. One interesting category is N-O-B. Here, robots deceive other robots or non-human objects to benefit the deceived other's benefits. As shown in the example, the robot sheepdog tries to deceive animals to believe it is an intentional dog or human using it behaviors [168]. Therefore, we can categorize it as N-O-B type. Similar to the robot sheepdog, robosquirrel also tries to deceive rattlesnakes using its behaviors [78]. By modeling the squirrel's behaviors to the mobile robot, it deceives rattlesnakes to believe this robot as a prey, and finally, rattlesnake behaviors can be observed for long-term studies.

Among these robot deception types, my research question is obviously related to H-O-B type, since it is aimed to find whether human can get any benefits from a robot's deceptive behaviors. Again, the H-O-B type is specified as a robot's deceptive behaviors to deceive a human partner for the deceived human's advantages. In sum, the research in this dissertation will find out how this type of deception can be applied to a robotic system and whether H-O-B robot deception is truly beneficial and applicable in appropriate HRI contexts.

3.3 Robot Deception: A case study

A taxonomy for robot deception was presented above by defining salient dimensions and categories. In this subsection, my previous research on robot deception is presented as an example and it is explained how my taxonomy fits to this specific work. In my previous research [139], a robot's deceptive behavior for resource protecting

Figure 5: Black Eastern Gray Squirrel moving peanuts
source by: `https://en.wikipedia.org/wiki/File:Eastern_Grey_Squirrel-black.jpg`

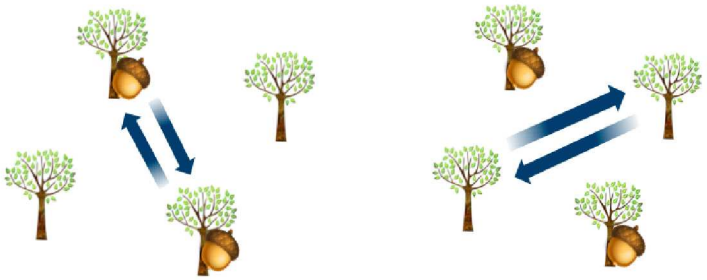
strategies, which is potentially applicable in military context and inspired by biology, was developed and evaluated.

3.3.1 Biological Findings

The patrolling strategy used by Eastern Grey Squirrels (Figure 5) is one interesting example in nature regarding the possible role of deception [155], where they use deception to protect their food caches from other predators. After hoarding food items, squirrels begin to protect their resources from pilfering by patrolling the caches. As the patrolling strategy, squirrels first move around the caching areas and check whether the cached food items are safe (Figure 6(a)). It was observed, however, that when a predator is present, the squirrel changes its patrolling behavior to spend time visiting empty cache sites, with the apparent intent to mislead the raider into the belief that those sources are where the valuables are located, a diversionary tactic of sorts (Figure 6(b)).

3.3.2 Computational Model and Implementation

Inspired by these deceptive behaviors of squirrels, a bio-inspired behavior-based model [12] of squirrel caching and protecting behaviors for application to robotic systems is developed and implemented in MissionLab, a mission specification software package developed by the Mobile Robotics laboratory at Georgia Tech [92]. MissionLab provides a graphical user interface that enables users to easily specify behavioral states



(a) True Patrolling (b) Deceptive Patrolling

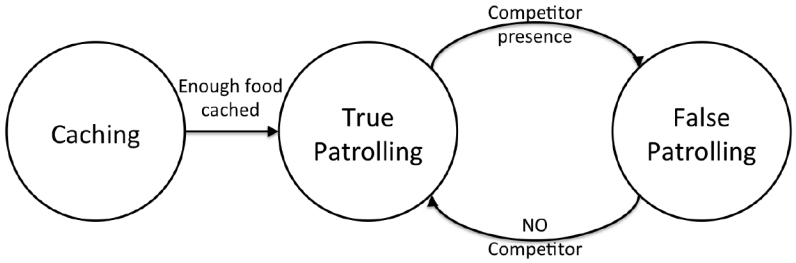
Figure 6: Squirrel's Cache Protection Strategy

and the control transitions between states, yielding a finite state acceptor (FSA), which can then be compiled down to executable code for both simulations and robots [91]. Each behavior component is an assemblage, a coordinated aggregation of primitive behaviors. The new caching and patrolling behaviors created are combined with pre-existing behaviors, such as avoiding obstacles, moving toward an object, or injecting randomness (noise). Simulation studies and real robot experiments were also performed to validate the algorithm.

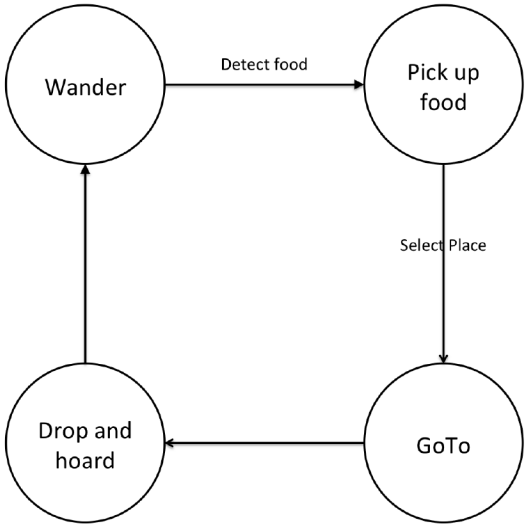
Figure 7(a) illustrates the high-level model of robot behaviors using a finite state acceptor (FSA). It starts from the caching behavior, but if any of the caching location is enough by the food items, it transitions to the patrolling strategy. In the patrolling strategy, if the competitor robot is nearby, the squirrel robot performs the deceptive patrolling strategy. Otherwise, the true patrolling strategy is procedure. Figure 7(b) and 7(c) illustrate the caching and the true/deceptive patrolling substrategies briefly. In the following subsections, each strategy will be explained with more implementation details.

3.3.2.1 True Patrolling Strategy

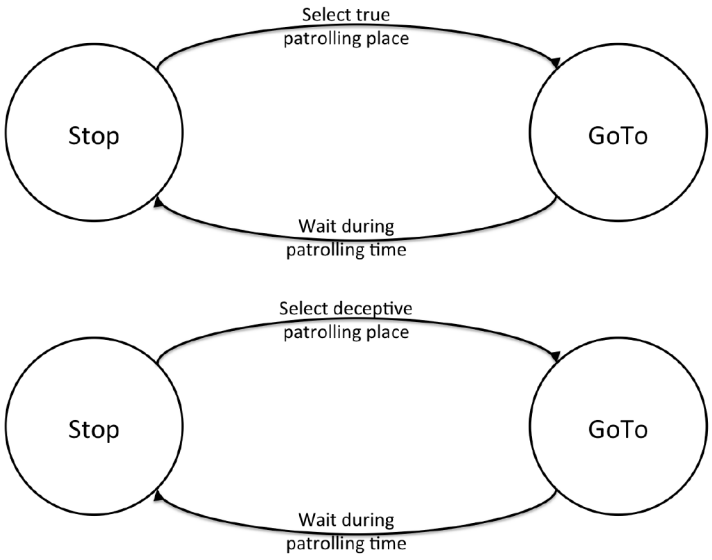
The robot's patrolling behaviors are developed and implemented in MissionLab. To implement the patrolling behaviors between the caching locations, goal-oriented movement, selecting places, and waiting behavior are used.



(a) High-level FSA: caching behaviors of squirrels



(b) sub-FSA: Caching



(c) sub-FSA: Food patrolling

Figure 7: Abstract level of Finite State Acceptors for squirrel robot's behaviors

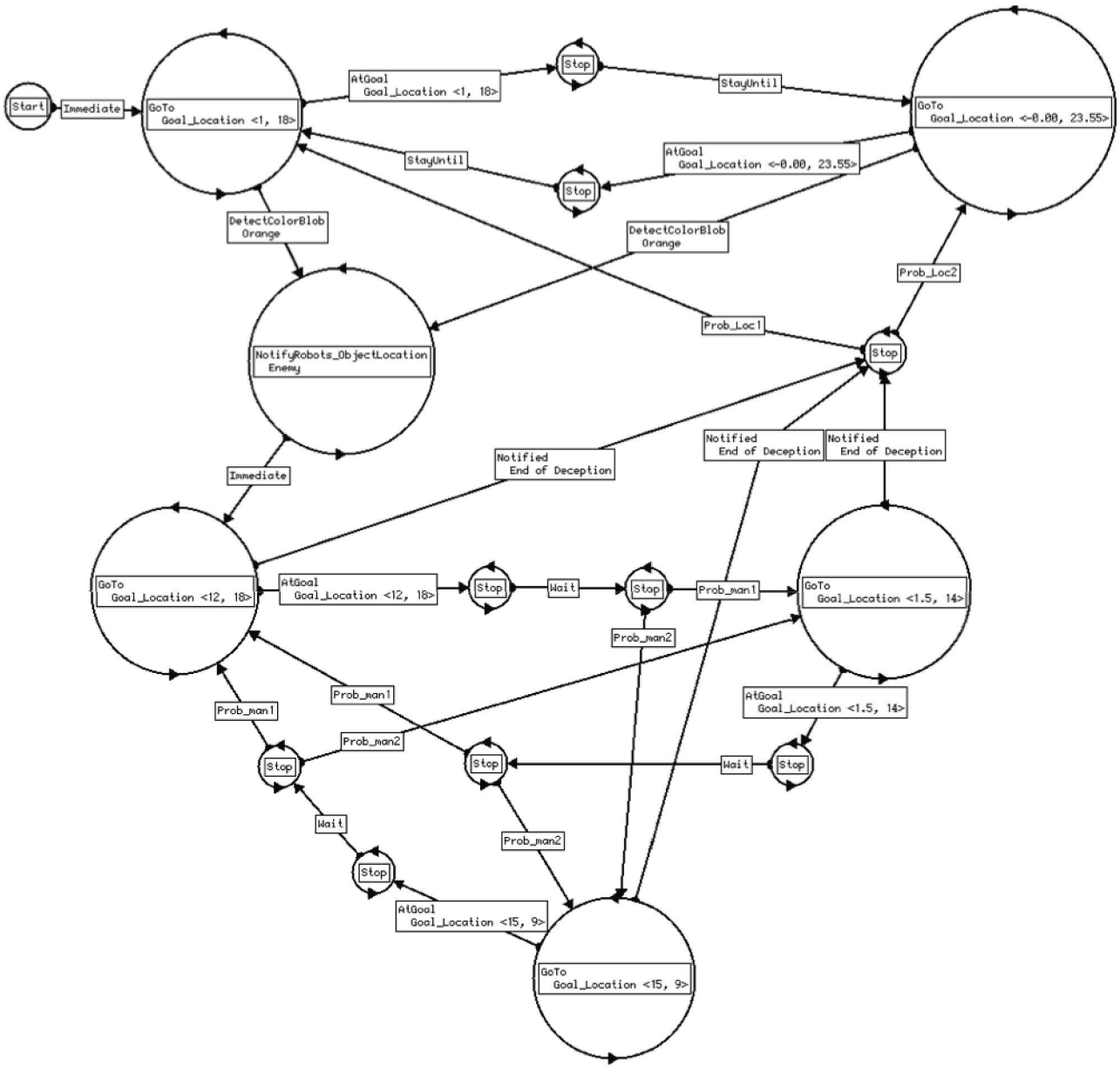


Figure 8: Finite State Acceptors for squirrel robot's patrolling strategy in MissionLab

Initially, the robot employs the true patrolling strategy by selecting one of the true patrolling locations. To select the patrolling location, the trigger calculates which of the many caching locations the robot should patrol. The calculation results in a random cache selection based on the transition probabilities among the places. The probabilistic transition model is used to determine the caching location. In patrolling, the transition probabilities are first determined by the number of previously cached items. In other words, if a place has more items, the probability that a robot will visit that location is higher. Therefore, the transition probabilities are calculated by the following equation:

$$P_{ij} = \frac{\#items_j}{\sum_{1 \leq k \leq n, k \neq j} \#items_k}$$

Here, P_{ij} is the transition probability that indicates that the location j is selected as the next patrol location when the current location is in location i. n is the total number of locations and $\#item_x$ indicates the number of food items in location x. The next patrol state is determined based on these transition probabilities. When the squirrel robot reaches a cache, it calculates the transition probabilities to other patrolling locations and decides on one of the transitions using a weighted roulette algorithm [71] (details can be found in Appendix). When a robot arrives the cache, it remains there for a finite amount of time similar to the patrolling behavior of an actual squirrel. In the true patrolling strategy, the time spent at the cache is determined by the number of food items in that place. At the end of the waiting phase, the robot selects the next patrolling location using the probability transition model discussed above and heads off to the next patrolling state.

Figure 8 illustrates the MissionLab FSA implementation of the squirrel robot's patrolling strategy. This FSA was implemented for the real robot experiments. Due to the limitations of lab space, I used two true caching locations and three deceptive caching locations. Therefore, in the true patrolling part, the robot moves back and

forth between two locations. When the robot detects the competitor robot, which has an orange marking, it transitions to the deceptive patrolling strategy. Here, I set three deceptive locations, and the transitions among these three locations are determined by the transition probabilities model that were explained above. Further details of deceptive patrolling strategy will be discussed in the next section.

3.3.2.2 Deceptive Patrolling Strategy

When the squirrel robot detects the presence of a competitor, deceptive behavior is triggered and the squirrel robot patrols the false (empty) caching locations to attempt to deceive the competitor. All objects and robot agents are marked by specific colors. Therefore, in our implementation, the deceptive patrolling strategy is activated by the DetectColorBlob trigger. In the deceptive patrolling strategy, the squirrel robot moves to and stays among the different deceptive caching locations. These locations actually include no food items, and the squirrel robot tries to mislead the competitor robot by visiting these false places. Again, the selection of deceptive locations is also calculated by transition probabilities. Here, the transition probabilities among the false locations are set as uniform distributions. In other words, the probabilities of each location are distributed equally.

As shown in Figure 8's deceptive patrolling part, the squirrel robot selects one deceptive caching location among several places based on transition probabilities. When the squirrel robot arrives in the deceptive caching location, it stays there for a while (time to stay is empirically set) to show the deceptive patrolling behavior to the competitor. After patrolling, it again determines the next deceptive patrolling location and repeats the patrolling behaviors. When the competitor robot is no longer detected in the vision of the squirrel robot, the end of deception trigger is activated and it returns to the true patrolling strategy.

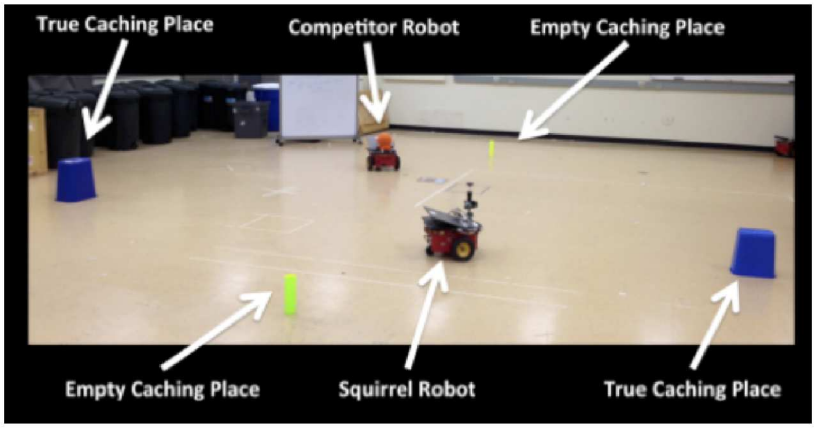


Figure 9: Robot Experiment Layout

Figure 10: Robot Platform: Pioneer robot with omni-directional camera

3.3.3 Experimental Results

This is a form of misdirection, where communication is done implicitly through a behavioral change by the deceiver. This strategy was implemented in simulation [139], and showed that these deceptive behaviors worked effectively, enabling robots to perform better using deception than without with respect to delaying the time of the discovery of the cache (see Appendix for more details). The real robot experiment was also performed using the experimental layout in Figure 9.

Two pioneer robots were used for the robot experiment: one as a squirrel robot and the other played the role of a competitor robot (Figure 10). The robots used an additional camera sensor for detecting another robot and the caching locations. The external omni-directional camera enabled the squirrel robot to observe 360 views of

(a) True Patrolling: GoTo Cache 1 → Patrol Cache 2 → GoTo Cache 2

(b) Competitor Detecting: Competitor Approaching → Detect Competitor → Change Behavior

(c) Deceptive Patrolling: Start Deceptive Behavior → GoTo Empty Cache → Patrol Empty Cache

Figure 11: Robot Experiment Scenario

the scene. Kumotek Robotics' omni-directional sensor was used with a Chameleon CCD camera[1].

Figure 11 illustrates the scenario of our robot experiment. During the experiment, the squirrel robot patrolled true caching locations to observe the resources. Since there were two caching locations in this simple scenario, the squirrel robot conducted back-and-forth movements in this states (Figure 11(a)). Next, the competitor robot started to wander around the area. At the end of scenario, the squirrel robot detected the competitor (Figure 11(b)), where it changed its true patrolling behaviors to deceptive patrolling behaviors. The squirrel robot started to move to the empty caching locations, and repeated these behaviors until the competitor robots left the area (Figure 11(c)).

To test the performance of the deception algorithm, it was measured how often the squirrel robot successfully deceived the competitor robot in this scenario. In the

[1]Omni-directional Sensor: Kumotek's VS-C450MR-TK model `http://www.kumotek.com/`
Chameleon CCD camera from Point Grey: `http://www.ptgrey.com/products/chameleon/`

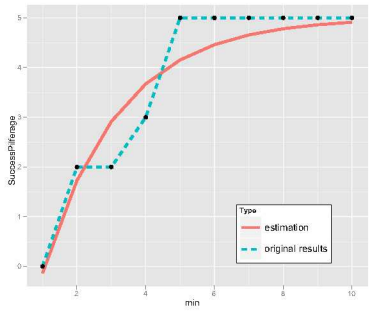
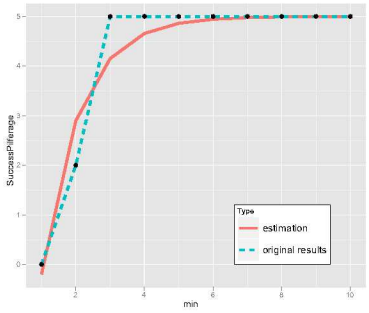
experiment, the squirrel robot and the competitor robot ran based on their FSA and behaviors as explained above. The competitor robot was declared to have successfully pilfered the cached items if it found the true caching locations within the specific time period, t. This experiment time t is one of the independent variables and it ranged from one to ten minutes. Each trial ran the experiment five times and it was observed how many times the competitor robot successfully pilfers the true locations for all 5 runs. Thus, the maximum successful pilfering can be five in each case. The experiments were performed under two different conditions; the squirrel robot with deceptive capabilities and the squirrel robot without deceptive capabilities (another independent variable).

As the time allowed increases, the number of successful pilferages also increases and converges to the maximum pilferage number, five. Based on this convergence rate, the performance of the algorithm can be evaluated. Faster convergence to the maximum pilferage number indicates that the algorithm enables the competitor robot to find the true caching locations more easily and more rapidly. In other words, slower convergence rates illustrate that the squirrel robot can protect resources longer and better.

To determine the convergence rate, the experimental results are plotted as shown in Table 5. By observing the plot, the estimation graph of each result can be formulated with the following equation, $y = \alpha + \beta \cdot e^{cx}$. This function is calculated using a non-linear least-squared regression method. The graphs in Table 5 show the results from the experimental data and their estimation functions. The green lines show the original experimental results, which are the number of successful pilferages (out of five) for each time period, t. The red lines indicate the regression functions for convergence.

In the exponential function, $y = \alpha + \beta \cdot e^{cx}$, the convergence rate depends on the exponentiation, parameter c. Simply, $1/c$ can determine the rate of convergence and

Table 5: Robot Experiment Results with Convergence Rate

	With Deception	**Without Deception**
Result		
Estimated Equation.	$y = 5 - 12.86e^{-0.9x}$	$y = 5 - 8.06e^{-0.45x}$
Convergence Rate	-2.2188	-1.1035

a larger value of the convergence rate indicates a faster convergence speed. As shown in the results from Table 5, the convergence rate of the "with deception" condition is -2.2188, which is smaller than the convergence rate under the "without deception" condition (-1.1036). Thus, it was observed that the squirrel robot using deception could protect the true caching locations longer than without deception capabilities. Therefore, it can be concluded that the deception algorithm leads to a robot's better resource protection performance.

3.3.4 Robot Deception Taxonomy

This previous research focuses on deceptive behaviors of robots in the military domain, where robots may hide and protect resources from other autonomous agents. The main purpose of this subsection is illustrating how robot deception taxonomy can be used with a real example. Obviously, this squirrel-like deception capability for a robot can be categorized in terms of my taxonomy. First, the object that a robot tries to deceive is other competitor robot, which means nonhuman objects (N). The deception happens through the robot's behaviors by intentionally misleading the competitor robot (B). In deception goal dimension, the benefits of this deception capability are

protecting the deceiver's resources longer, so the squirrel robot obtains advantage: i.e., self-oriented deception (S). As a result, this squirrel robot deception is classified as N-S-B type in our taxonomy.

3.4 Summary

Deception is one of the capabilities that is needed to achieve higher-order intentionality. Therefore, deceptive capabilities are desired to add to robot systems, especially focusing on social robots. However, there is a lack of studies on fundamental theory, such as the definition of a taxonomy for robot deception. As a preliminary work, previous research on deception was reviewed, and a novel taxonomy for classification of robot deception was developed in my research. Also, by presenting my previous squirrel-like robot deception work, it was illustrated how this taxonomy can clearly apply for the real work example.

Again, among different robot deception types, this dissertation's research question is focusing on to H-O-B and H-O-P types, since the purpose of this research is to find whether human can get any benefits from a robot's deceptive behaviors. Especially, it is essential for social robots to perform the deceptive behaviors only when the deceived humans' benefits are expected. In other words, the goal of deception should be a key factor developing robot deception in HRI contexts. In sum, this research aims to determine how other-oriented deception can be applied to a robotic system and whether it is truly beneficial and applicable in appropriate HRI contexts.

CHAPTER IV

COMPUTATIONAL ARCHITECTURE

In the previous two chapters, basic knowledge related to deception and developed fundamental theories for my robot deception work was presented. From the literature reviews in chapter 2, several clues on why deceptive behaviors are essential in a robotic system were found. More importantly, the use of other-oriented deception has been extensively reviewed, and a robot's other-oriented deception has also been defined in chapter 3. Based on these fundamental theories, this research aims to investigate the use of other-oriented robot deception in HRI contexts and eventually, to achieve a more socially intelligent and intentional agent.

As stated in this dissertation's introduction, the research in this dissertation argues that the robot should always be beneficial to humans in HRI. Therefore, robot deception should only be used when it can provide advantages to deceived humans. In other words, it is necessary to develop and test other-oriented robot deception in the context of HRI.

To achieve other-oriented robot deception, a computational model for a robot's other-oriented deception should first be developed and implemented into an appropriated robot platform. In this chapter, a novel computational model for a robot's other-oriented deception will be presented. After successful implementation, the research hypothesis needs to be tested and proved via appropriate HRI studies, which will be discussed in the following chapter (chapter 5). Applying deception capabilities to the robotic system also leads to multiple ethical issues when particularly discussing social robots. Therefore, these ethical issues will also be discussed again in the later chapter (chapter 6).

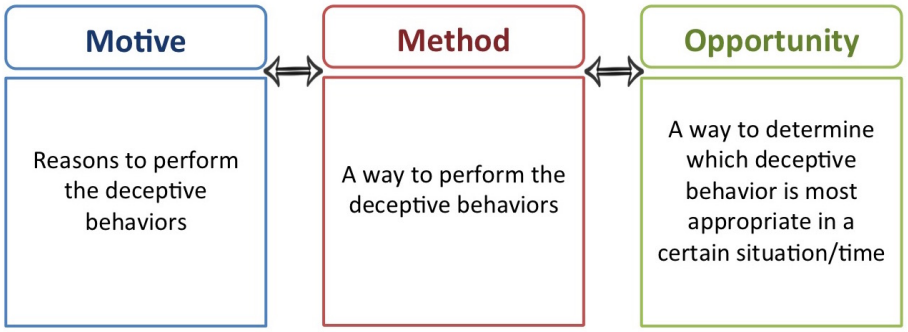


Figure 12: Criminology-inspired computational architecture for a robot's other-oriented deception

4.1 Computational model for a robot's other-oriented deception

Inspired by a criminological definition of deception [5], a computational model of a robot's other-oriented deception is developed in my research [141, 142]. According to criminological findings, deception is analyzed by three criteria, which are **motives**, **methods**, and **opportunity** (Figure 12). **Motives** refer to the reasons for deception. **Methods** discuss how to perform the deceptive behaviors, and **opportunity** refers to when the deceptive behavior is most appropriate. In previous work, the deception model included two dimensions, which are *when* and *how* criteria [176]. However, my model differs from this previous model because the goal of this model is for a robot to have the capability to only perform other-oriented deception. In other words, besides when and how, it is also essential to discuss the *why* problem in this model. From my taxonomy (in Section 3), the interaction goals, which are self-oriented deception and other-oriented deception, should be reflected in the computational model. In the criminological approach, motive can be discussed with this dimension, and therefore, the development of my model for robot deception is inspired by this approach.

Based on this criminological law, an algorithm of robot deception is developed. In a high-level view, it is first necessary to determine whether the current HRI context includes any motives for a robot to perform the deceptive behaviors. If so, then

a robot should generate the methods to perform deception, which are alternative deceptive behaviors beyond the normal true action(s). Finally, by selecting among different true/deceptive behaviors, it should be possible to determine which one is the most appropriate in a certain situation, thus providing opportunity. According to this approach, an algorithm for each criterion is separately developed and then integrated together to achieve the computational models for a robot's other-oriented deception.

4.2 Method: Deceptive Action Generation Mechanism

Methods (means) define the way in which the deception is performed. It is necessary to build a model that illustrates how deceptive actions can be generated, where we aim to determine the set of true/deceptive actions that a robot performs during the interaction. For this, deceptive action generation mechanism has been developed in my model [141].

A human behavior is manipulated by verbal and non-verbal actions. When a robot delivers information to humans and interacts with them, the robot uses several cues for representing the action. For verbal delivery a robot uses multiple verbal cues, including speech expressions and vocal tones [123]. Non-verbal communication actions involve the robot's bodily cues, which include gesture, facial expression, and proximity [22]. A robot's action of this sort can be formulated as $A =< a_v, a_n >$, which indicates the combination of verbal action a_v and nonverbal action a_n.

A robot's deceptive action in this model is focused on non-verbal communication display behaviors a_n. By generating the information using bodily cues, humanoid robots can reap certain advantages [26]. First, nonverbal actions often have benefits that transcend cultural norms [26]. In HRI contexts, a robot is limited in its verbal interactions due to language differences. However, humans can interpret nonverbal expressions somewhat more generally. In addition, people may expect a humanoid

robot to demonstrate nonverbal actions due to its embodiment. Therefore, these bodily expressions can lead to more natural interactions between humans and robots. Finally, nonverbal actions potentially increase the probability of forming bonds of trust and affect between humans and robots [26, 84].

Due to these advantages of nonverbal actions, a set of a robot's true/deceptive actions is defined using nonverbal cues. In a high-level view, to generate a robot's deceptive actions, a robot should first have a default action, which is a true action a_t. Then, according to the deception generation mechanism described below, the robot can generate a set of deceptive actions by transforming the selected default true action. A robot can also have multiple true actions that can be applicable to the current situation. Therefore, the set of true actions can be defined such as $A_t = \{a_{t1}, a_{t2}, \ldots a_{tn}\}$.

4.2.1 Deception Generation Mechanism

According to Bell and Whaley [18], deception can be categorized into two main types - hiding and showing (Table 6). Type 1 deception is hiding, which means masking characteristics of the truth to represent deception. Type 2 deception is showing; it aims to deceive the mark by representing false information. The deception generation is modeled based on this categorization. In other words, a robot generates deceptive behaviors by transforming the default true action consistent with these two deception mechanisms.

4.2.2 Generating Deceptive Action

As stated above, this model intends to generate a robot's deceptive action using nonverbal behaviors. This nonverbal action is represented by several bodily cues, including body gestures (g), facial expression (f) and proximity (p). Therefore, a robot's action can be formulated as $a =< g, f, p >$. As shown in this formulation, the nonverbal action a is generated by combining these three different cues, but not all

Table 6: Deception Generation Types

	Mechanism	**Explanation**
Type 1	Deception by Omission (DbO)	Hiding information; the true action will be transformed by deleting key information
Type 2	Deception by Commission (DbC)	Showing false information; if changeable key information exist, the action will be transformed by changing the values of these key information

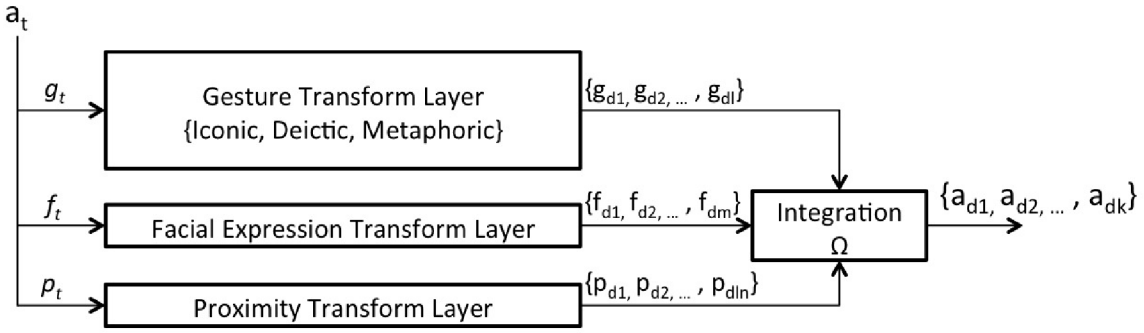


Figure 13: Overview of the action generation mechanism via deception transformation layers for nonverbal action cues

cues need to be included every time. These bodily cues are manipulated differently to generate the deceptive actions in each cue. The means by which these transformations occur are described below.

After the default true action is selected for a robot system, the deceptive actions are then generated. Again, the true action is a combination of bodily cues $< g, f, p >$. Each cue is transformed to its deceptive action form(s) separately during deceptive action generation. As shown in Figure 13, each action cue inputs to the deception generation layers, and when the deceptive action cues are generated, these cues are combined together to construct the deceptive actions $a_{d1}, a_{d2}, \ldots, a_{dn}$. The way to generate deceptive action cues in each layer is varied, and the mechanisms for each bodily cue are explained below.

4.2.2.1 Gesture Transform Layer (g)

Previous research in nonverbal behavior has divided a robot's body gestures into four categories [23]:

1. Iconic gesture (g_{iconic}): meaningful motions associated with the semantic content of speech.

2. Deictic gesture ($g_{deictic}$): motions to guide attention toward specific objects in the environment. This type of gesture is generally prototyped by pointing actions.

3. Metaphoric gesture ($g_{metaphoric}$): motions to represent abstract concepts; behavioral fragments that convey implicit information without being tied to dialog.

4. Beat gesture (g_{beat}): simple up-and-down movements to emphasize certain words or phases.

Among these four gestures, semantically meaningful actions without speech can be found in iconic, deictic, and metaphoric gestures. Therefore, beat gesture is excluded in this deceptive action generation model. In other words, a robot's gesture cue g is defined by one of three action types (iconic, deictic, or metaphoric gestures), and deceptive gestures with semantics can be generated by the manipulation of these three categories as described below.

Iconic gestures are gestural representations of the semantics of spoken language in general. Therefore, the transformation of iconic gestures depends on the information that a robot wants to deliver to the human via speech. To represent meaningful information, humans generally use hand gestures. For example, a specific number can be shown using fingers. We can also define a robot's iconic gestures based on meaningful hand and arm gestures. When the robot has a true default hand gesture, deceptive gestures can be created according to the two deception types as shown in

table 6. First, it can hide the information by simply not displaying any iconic gestures (deception by omission). In deception by commission, a robot can change the information displayed in the true gesture by giving variations. For example, assume that a robot's true action is showing the number three with its fingers. In this case, this finger representation illustrates a semantically meaningful number, so it is an iconic gesture. Here, for type 1 deception (omission), a robot can just not show any hand gestures to the human. In type 2 deception (commission), a robot's finger signaling gesture can be varied to other numbers such as one or two.

Deictic gestures also include important information that is useful to transfer to humans. Archetypal deictic gestures include pointing actions; therefore, a transformed deceptive action can be determined by changing the direction of pointing (Type 2 - commission) or not pointing at all (Type 1 - omission). A rotation of the head and torso is often associated with the arm pointing gesture. For example, the default deictic action is to point in the direction of a specific object, whereas the deceptive deictic gesture can be generated by shifting the direction of pointing toward other objects or other spaces.

Metaphoric gestures represent abstract concepts. Particularly, humans can express and deliver their emotional status via gesture, and these emotional expressions are categorized as metaphoric gestures in general. Therefore, emotional gestures can be added as the metaphoric category to the robot system. Human emotion can be classified into six categories, which contain happiness, anger, fear, surprise, disgust, and sadness [52]. In addition, neutral emotion can be included where the robot has no metaphoric expression. Based on these seven categories, a set of default expressions for a robot can be defined. For deceptive action generation, when a robot selects the true emotional gesture, it can determine deceptive metaphoric gestures by selecting an opposing emotional expression (Type 2 - commission) or by not showing any emotion using a neutral gesture (Type 1 - omission).

Table 7: General Gesture Primitives (notation: ggp_i) with necessary parameters and body parts in a humanoid robot; ggp_2 is the deictic gesture and all other gesture primitives are iconic gestures.

General Gesture (notation) [parameter]	**Body Part**
Idle (ggp_1)	Head, Left and Right Arms, Legs
Raising/Showing Hand (ggp_2) [# of fingers]	Right Arm
Hiding Hand (ggp_3)	Right Arm
Grasping (ggp_4) [object Location]	Head, Right Arms, Legs
Pointing (ggp_5) [object Location]	Head, Right Arms, Legs
Waving (ggp_6)	Right Arms
Okay/Yes (ggp_7)	Head, Right Arm
No (ggp_8)	Head, Right Arm

Table 8: Emotional Gesture Primitives (notation: egp_i) with necessary parameters and body parts in a humanoid robot; These emotional gesture primitives are the metaphoric gestures.

Emotional Gesture (notation)	**Body Part**
Happiness (egp_1)	Head, Left and Right Arms, Legs
Anger (egp_2)	Head, Left and Right Arms, Legs
Fear (egp_3)	Head, Left and Right Arms, Legs
Surprise (egp_4)	Head, Left and Right Arms, Legs
Disgust (egp_5)	Head, Left and Right Arms, Legs
Sadness (egp_6)	Head, Left and Right Arms, Legs
Neutral (egp_7)	Head, Left and Right Arms, Legs

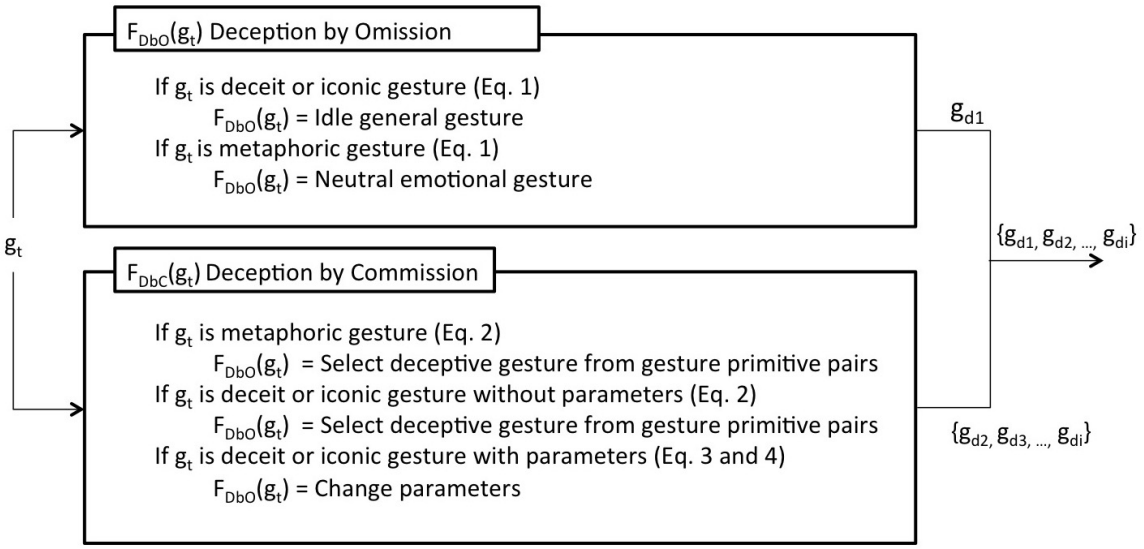


Figure 14: Overview of the action generation mechanism via gesture transformation layer

The robot's default (true) gesture can be generated from one or more of these three gesture main categories. In the current robot system and without loss of generality, robot gestures are generated by selecting/combining gesture primitives; eight general gesture primitives and seven emotional gesture primitives are defined as shown in Table 7 and Table 8 [159]. Gestures g_{iconic} and $g_{deictic}$ are produced by combining the general gesture primitives, and the metaphoric gesture $g_{metaphoric}$ is determined by selecting one of the seven emotional gesture primitives.

Now, it is necessary to define the deception generation function F for each gesture primitive. As stated above, deceptive gestures are generated by two types of deception (table 6), which are deception by omission (F_{DbO}) and deception by commission (F_{DbC}). Figure 14 illustrates the overview of deceptive gesture generation mechanisms. First, according to the deception by omission mechanism, a robot can perform a deceptive gesture by simply not showing the current gesture. In other words, as shown in Equation 1, when the robot has a true gesture primitive in any category, the robot can perform the deception by omission by changing it to the Idle

(ggp_1) / Neutral (egp_7) gesture primitive to realize the omission deceptive gesture set.

$$\begin{aligned} F_{DbO}(ggp_2|ggp_3|ggp_4|ggp_5|ggp_6|ggp_7|ggp_8) = ggp_1 \\ F_{DbO}(peg_1|egp_2|egp_3|egp_4|egp_5|egp_6) = egp_7 \end{aligned} \tag{1}$$

To generate a deceptive gesture according to deception by commission, the model needs a way to produce false information for each gesture primitive. Two means of generating false information are used in the system.

First, according to the characteristics of the gesture primitives, primitive pairs that contain gestures of opposite meanings are predefined, whereby the deceptive gesture can be determined by finding the opposite of each primitive gesture. For the general primitives, opposite pairs are defined as people recognize in general [159]. In addition, for the emotional primitives, these opposite emotion pairs are discriminated according to Plutchik's wheel of emotions [116]. As a result, the set of opposite gesture primitive pairs can be obtained as shown in Equation 2, which represents the mathematical formulation of the deception by commission mechanism. As shown here, the set of gesture primitive pairs is defined, and the robot can determine the opposite gestures based on this pair set P.

$$\begin{aligned} SetofGesturePrimitivePairs(P) = \{&[ggp_2, ggp_3], [ggp_7, ggp_8], [egp_1, egp_6], \\ &[egp_1, egp_5], [egp_2, egp_3]\} \end{aligned} \tag{2}$$

If $[g_1, g_2] \in P$, then $F_{DbC}(g_1) = g_2$ or $F_{DbC}(g_2) = g_1$

Figure 15 shows exemplar pairs of gesture primitives. All gesture primitives were implemented using the Webots simulator [36] and Choregraphe and with the NAO robot [151]. Figures 15(a) shows the "showing hand (ggp_2)" and the "hiding hand (ggp_3)" gesture primitives. Since these two gestures are in the set of gesture primitives

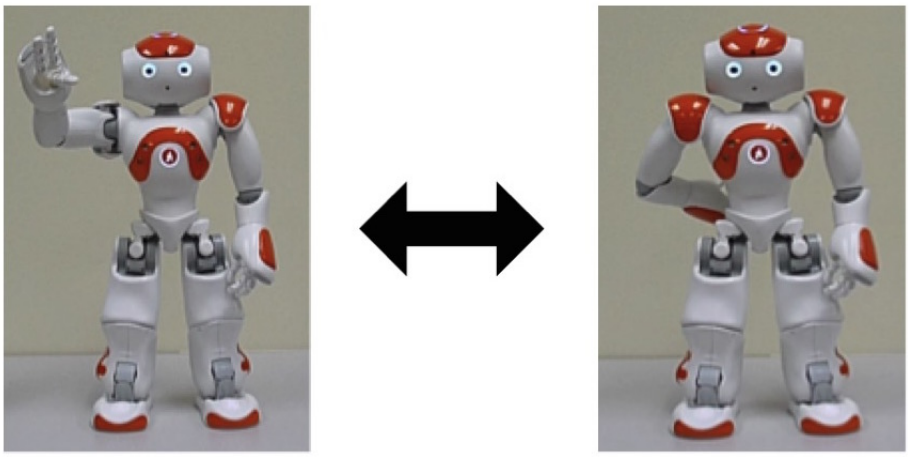
pairs P, when one of two gestures is selected as a true set, the alternate gesture is used as a deceptive gesture according to Equation 2. Figures 15(b) illustrates the emotional gesture pairs such as [$Anger(epg_2)$, $Fear(egp_3)$]. As shown in the final example, when "happy (epg_1)" gesture primitive is selected, "sad (epg_6)" or "disgust (epg_5)" gestures are selected as deceptive actions as shown in Figure 15(c).

Second, when the primitive gesture has a parameter that represents key information for the action, the deceptive gesture can be generated by changing this key value. Thus, if the value of the parameters are changed to different values, false information can be delivered to the mark, and, as a result, a deceptive gesture can be generated (commission).

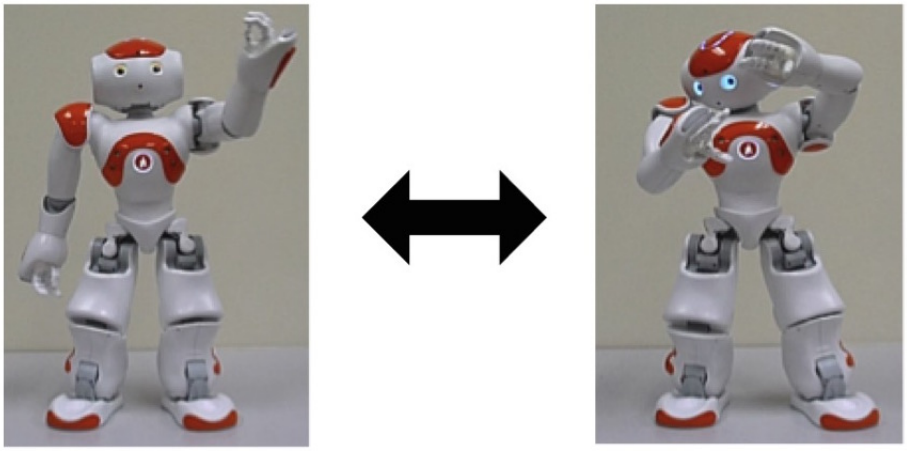
As shown in Table 7, ggp_2, ggp_4, and ggp_5 require a parameter to express gestures, and each primitive can be defined as $ggp_2(n)$, $ggp_4(x)$, and $ggp_5(x)$, where n and x specify the values of the parameters. Here, n represents the number of robot fingers and x is the directional vector of the intended object's location. For these three gesture primitives, the robot should generate the deceptive action by changing the parameter value to a false one as shown in Equation 3 and 4.

$$\begin{gathered} F_{DbC}(ggp_2(n_k)) = \{ggp_2(n_i)|n_i \in \{n_1, \ldots, n_{k-1}, n_{k+1}, \ldots, n_l\}\} \\ \text{where } n = \text{number of robot fingers}, 0 \leq n \leq n_l, \\ n_l = \text{max number of a robot finger} \end{gathered} \tag{3}$$

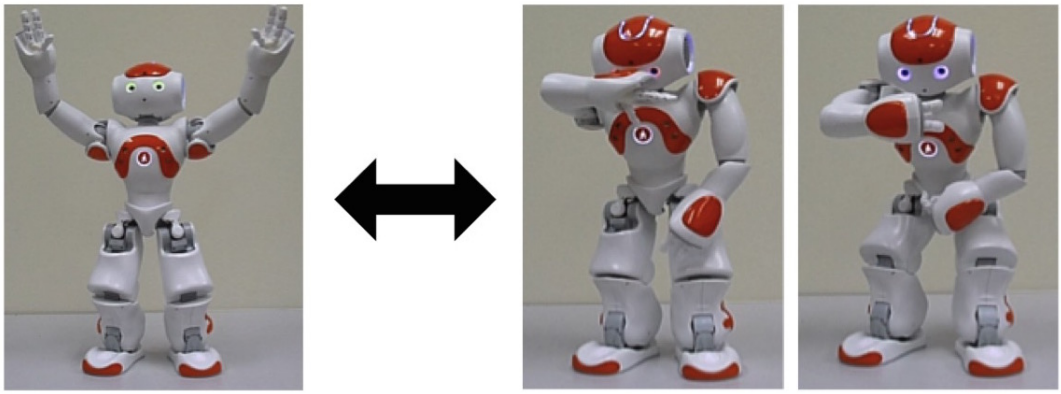
$$\begin{gathered} F_{DbC}(ggp_4(x_k)) = \{ggp_4(x_i)|x_i \in \{x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_l\}\} \\ F_{DbC}(ggp_5(x_k)) = \{ggp_5(x_i)|x_i \in \{x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_l\}\} \\ \text{where } x_i =< x, y, z > : \text{ vector of object's location} \end{gathered} \tag{4}$$



(a) Left:ggp_2 (showing hand) vs. Right:ggp_3 (hiding hand); $[ggp_2, ggp_3] \in P$

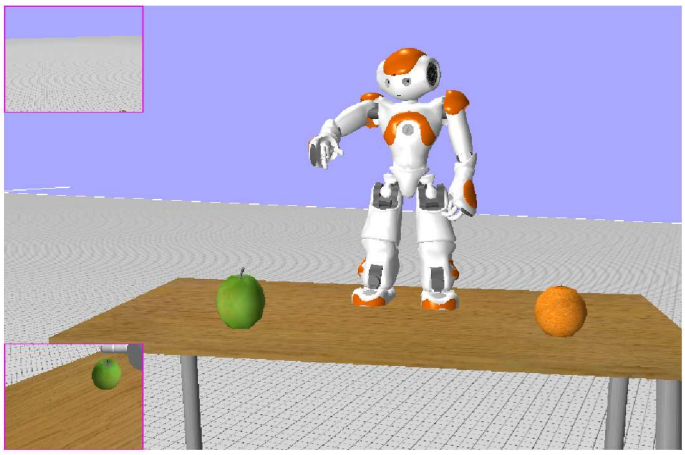


(b) egp_2 (anger) vs. egp_3 (fear); $[egp_2, egp_3] \in P$

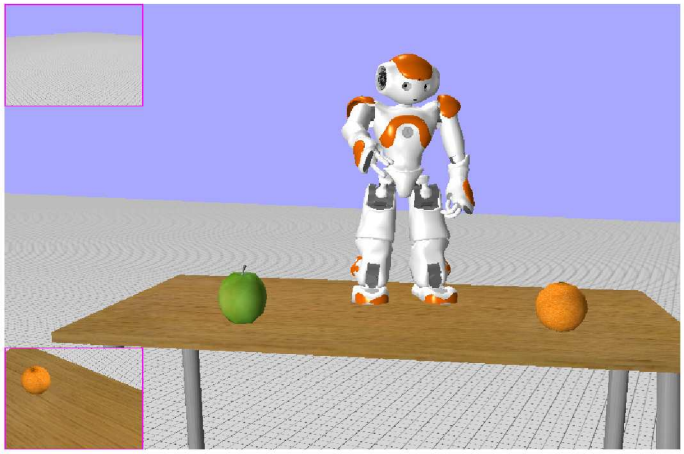


(c) egp_1 (happy) vs. egp_5 (sad) and egp_6 (disgust); $[egp_1, egp_5] \in P$ and $[egp_1, egp_6] \in P$

Figure 15: Examples of gesture transformation via Deception by Commission mechanism; Gesture primitive pairs which are in the set of gesture primitives pairs P. Therefore, when one of gestures is selected as a true set, the alternate gesture is used as a deceptive gesture according to Equation 2.



(a) True pointing action $ggp_5(apple)$



(b) Deceptive pointing action $ggp_5(orange)$

Figure 16: Simulations of deceptive "pointing" gesture generation via Deception by Commission mechanism. According to Equation 3, the alternative object's location is selected as deceptive pointing action position.

Figure 16 illustrates deception generation example via this deception by commission mechanism. The gesture primitive in this simulation is "pointing" gesture (ggp_5). In this simulation context, a robot detects two object locations {apple, orange}. When $ggp_5(apple)$ is selected as a true pointing action as shown in Figure 16(a), a robot can generate the deceptive pointing action $ggp_5(orange)$ based on equation 3 as shown in Figure 16(b).

In sum, by applying the deception by omission and deception by commission gesture generation functions, a robot can find alternative gestures that can be used to deceive the human. These principles can be generalized even further as needed.

4.2.2.2 Facial Expression Transform Layer (f)

A facial expression (human or robot) is usually used to display emotional states. Figure 17 illustrates the overview of the facial expression transform layer. As stated earlier, according to Ekman [52], emotion can be divided into six basic categories, which are happiness, anger, disgust, fear, sadness, and surprise. In addition, neutral status is commonly added to the emotion categorization. From a higher-level perspective, these facial expressions can fall into three sets —positive (f_p), negative (f_n), and neutral (f_{nt}). Positive facial expressions are a representation of happiness.

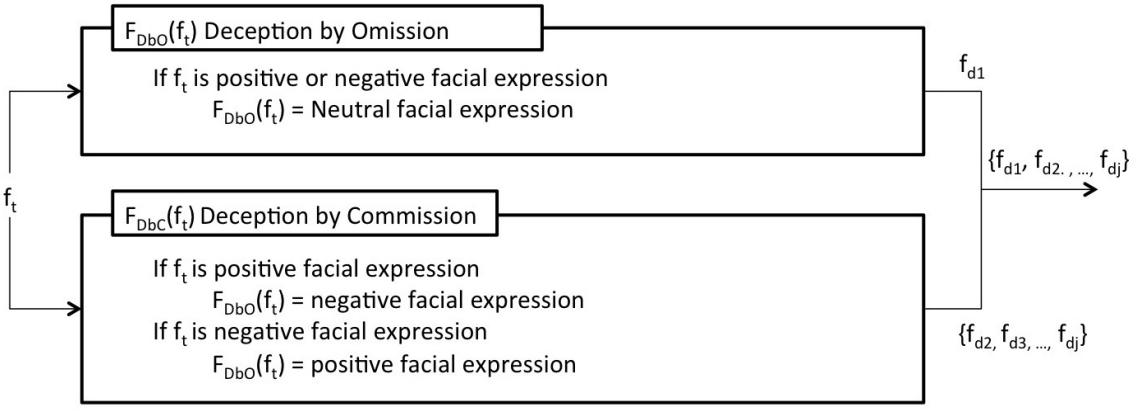
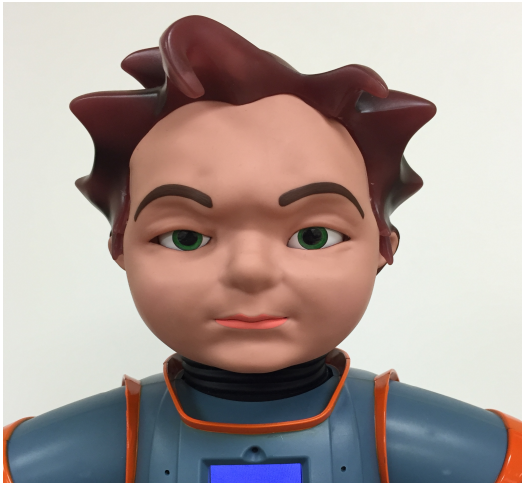


Figure 17: Overview of the action generation mechanism via facial expression transformation layer

(a) Default true facial expression: Positive



(b) Deception by Commission: $f_t \in \{f_p\} \rightarrow$ the deceptive facial expression f_d is transformed by selection from the negative or neutral facial expressions ($f_d \in \{f_n, f_{nt}\}$)

(c) Deception by Omission: The true action is to display the robot's emotional state $\rightarrow$ Not display any emotion, neutral facial expression f_{nt}

Figure 18: Example of deceptive facial cue generation with the R25 robot [126]

Negative facial expressions include all expressions of anger, disgust, fear and sadness. Neutral facial expressions (f_{nt}) are shown when a robot doesn't express any emotion. In this model, when a robot generates deceptive facial expressions, these three sets are used to determine the correct one to provide. It is first determined whether the true default expression is in the positive, negative, or neutral set. The robot can then transform the true facial cue by applying deception by commission. In other words, to show the false interaction, a robot selects from the other two orthogonal sets for an emotional display choice.

For example, as shown in Figure 18 if the default true facial expression f_t is positive ($f_t \in \{f_p\}$), then the deceptive facial expression f_d will be transformed by selection from the negative and neutral facial expressions ($f_d \in \{f_n, f_{nt}\}$). Here, this example is demonstrated with the R25 humanoid robot [126], because it can represent the facial expression effectively. When the robot does not have capabilities for such expressive facial representation, other indirect methods can also be used as an emotional facial expression cue (e.g., the color of NAO's ear/eye LEDs [104]).

4.2.2.3 Proximity Transform Layer(p)

Figure 19 illustrates the overview of the proximity transformation layer. Omission deception for facial expression is straightforward. If the true action is to display the robot's emotional state requiring such a display it will either not display any emotion

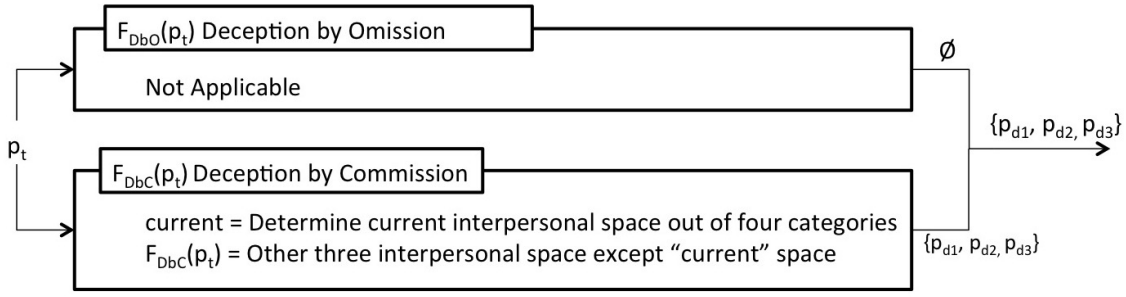


Figure 19: Overview of the action generation mechanism via proximity transformation layer

Table 9: Humanoid Robot's Proxemic Spatial Regions

Space Category	**Proxemics Zones**
Intimate, p_{in}	0-60 cm
Personal, p_{ps}	75-120 cm
Social, p_{sc}	150-200 cm
Public, p_{pb}	Over 200 cm

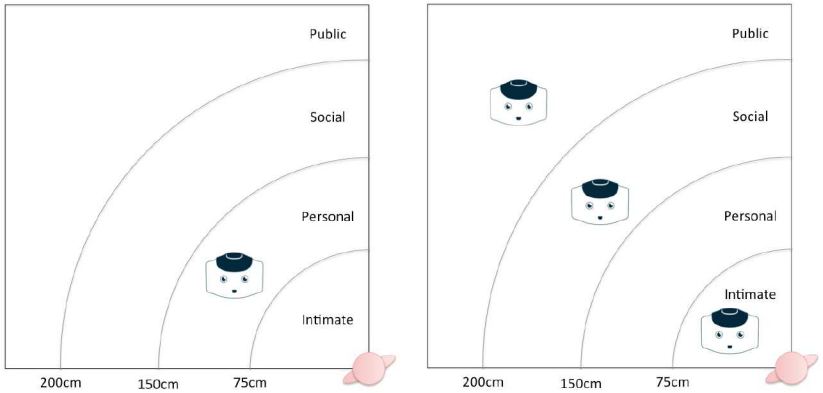
whatsoever, or if it is already displaying an emotional facial expression that should be changed according to the new true action, it will instead continue to display its previous facial expression without change.

Spatial proximity is indirectly used to give an impression of intimacy to humans during the interactions. Hall [64] divided interpersonal space into four categories: intimate (within 2 feet of the person), personal (2-4 feet), social (4-12 feet), and public (12-25 feet) spaces. Previous robotics research [26] has studied how these interpersonal spaces can be applied in HRI contexts by quantizing these four spaces separating human and robot as shown in Table 9. Therefore, a robot's proximity cue can be defined as a member of one of these four categories. This indicates the degree of familiarity with the human partner. For deception generation, the algorithm is developed similarly to facial expression mechanism (overview: Figure 19). When the default proximity cue lies in one of the four space categories, the alternative deceptive action set can be created by selecting from the other three space categories.

For example for type 2 (commission), if the default proximity is defined as personal space ($p_t \in \{p_{ps}\}$), the deceptive proximity set will be $p_d \in \{p_{in}, p_{sc}, p_{pb}\}$ as shown in Figure 20. For type 1 (omission), the robot should remain in its place even if the true action warrants a change in spatial separation.

4.2.2.4 Integration of deceptive non-verbal action cues

The previous subsections have explained a robot's deceptive action generation for each bodily cue type. Via these transformation layers, a robot can produce multiple



(a) True proximity cue (b) Deceptive proximity cue

Figure 20: Example for Type 2 (commission) deceptive proximity generation

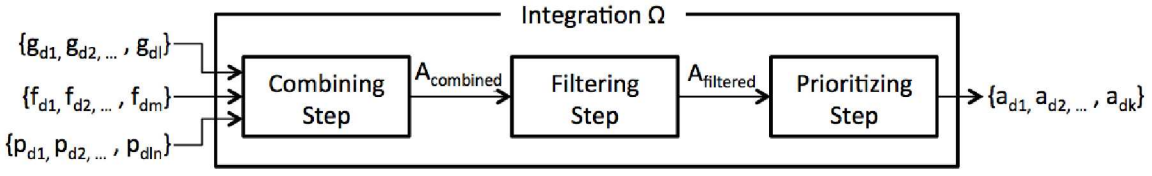


Figure 21: Detailed integration step of the action generation mechanism (extended from Figure 13)

deceptive actions. The final step in generating deception is integrating these discrete cues into one holistic robot action. As shown in Figure 13, this final step is defined as integration (Ω). The potential deceptive action $a_d =< g_d, f_d, p_d >$ is generated by combining the 3 elements of deceptive non-verbal action cues.

As shown in Figure 21, the integration module is structured in three steps: combining, filtering, and prioritizing. The pseudo-code for this integration module is described in Algorithm 1. As illustrated in this algorithm, a robot first generates all combinations of possible deceptive bodily, facial expression, and proximity cues and gets the set of possible deceptive actions such as $A_{combined}$. The robot can easily obtain the set of possible deceptive actions by generating all combinations of deceptive bodily cues.

From the set of possible deceptive actions, some of the actions should be rejected due to potential contradictions. For example, if the facial expression cue shows the positive emotion but the gesture cue delivers the sadness motion, it will lead to confusion in the human subject. To avoid those contradictory actions, a filtering step is added here. In the filtering step, a robot checks whether the current action's bodily and facial expression cues are globally coordinated as shown in Algorithm 1.

Again, the contradiction can potentially occur when each action cue in one action tuple shows extremely different information at the same time. As described in Section 4.2.2.1, robot gestures can be categorized in three ways: iconic, deictic, and metaphoric. General gesture primitives are used to represent iconic and deictic gestures and metaphoric gestures can be produced by emotional gesture primitives.

Facial expression cue is used to show the emotional state of the robot; therefore, it is potentially overlapped with the metaphoric dimension in the gesture cue. Therefore, a check for potential conflict between emotional gesture cue and facial expression cue is made. Since these two cues express emotional state concurrently, the contradiction can occur if two cues show extremely different motions. Therefore, when a negative

Algorithm 1 Integration of deceptive action cues

Input: Deceptive non-verbal action cues from three transform layers
$G_d = \{g_{d1}, g_{d2}, \ldots, g_{dl}\}$, $F_d = \{f_{d1}, f_{d2}, \ldots, f_{dm}\}$, $P_d = \{p_{d1}, p_{d2}, \ldots, p_{dn}\}$
Output: Deceptive Action Set $A_d = \{a_{d1}, a_{d2}, \ldots, a_{dk}\}$

1: *// **Step 1.** Combining Step*
2: *// Generate all possible deceptive actions by combining deceptive gestures, facial expressions, proximities*
3: $A_{combined} = \{< g_d, f_d, p_d > | g_d \in G_d, f_d \in F_d, p_d \in P_d\}$
4: *// $A_{combined}$ is a variable to store all possible deceptive actions*
5:
6: *// **Step 2** Filtering Step*
7: *// Set the high-level emotional primitive gesture group (positive, negative, neutral)*
8: $E_{positive} = \{egp_1, egp_4\}$ *// epg_1 = happiness, egp_4 = surprise*
9: $E_{negative} = \{egp_2, egp_3, egp_5, egp_6\}$ *// egp_2 = anger, egp_3 = fear, egp_4 = disgust, egp_5 = sadness*
10: $E_{neutral} = \{egp_7\}$ *// egp_7 = neutral*
11:
12: *// Find contradictory emotional cues and remove them*
13: $A_{filtered} = \{\}$ *// initialize variable $A_{filtered}$, a variable to store the actions after removing any actions that contain contradictions*
14: **for** each action tuple $< g_d, f_d, p_d >$ in $A_{combined}$ **do**
15: **if** $!(g_{di} \in E_{positive}$ && $f_{di} \in f_n)$ && $!(g_{di} \in E_{negative}$ && $f_{di} \in f_p)$ **then**
16: *// metaphoric gesture and facial expression pairs that represent opposite emotion are defined as contradiction/conflict*
17: $A_{filtered} = A_{filtered} \cup < g_{di}, f_{di}, p_{di} >$ *// non-contradictory action tuples are only stored in $A_{filtered}$*
18: **end if**
19: **end for**
20:
21: *// **Step 3.** Prioritizing Step*
22: *// Avoid conflict of actuators by performing bodily cues that possibly use the same joints/motors in different times*
23: t_{start} = time to start deceptive action cue *// to perform each action cue in different times, set time variable t_{start}*
24: $t_{proximity}$ = time duration to complete the proximity cue
25: **for** each action tuple $< g_d, f_d, p_d >$ in $A_{filtered}$ **do**
26: $t_1 = t_{start} + t_{proximity}$ *// time to start gesture primitive cue - after performing proximity cue*
27: $t_2 = t_3 = t_{start}$ *// facial expression cue can be performed with proximity cue at the same time*
28: $A_{di} =< g_{d_i}^{t_1}, f_{d_i}^{t_2}, p_{d_i}^{t_3} >$ *// add start time variations to action cues*
29: $A_d = A_d \cup \{a_{di}\}$ *// add to final deceptive action set*
30: **end for**

---

emotion gesture and a positive facial expression are shown in the same action a_i, it should be filtered out. The same step occurs in the case of an action with a positive emotion gesture and negative facial expression. As a result, in our algorithm, the sets of positive, negative, and neutral emotional primitive gestures are defined first based on Plutchik's definition [116]. Then, it is determined whether the facial expression cue is in a contradictory emotional group, and, when those two cues are not in the same emotional group, it is removed.

Proximity is highly related to the intimacy and it can indirectly deliver the emotions to human subjects [64, 162]. Therefore, proximity is also aligned with the group of metaphoric gestures. However, it is difficult to determine the specific type of emotion that the proximity affects. Therefore, proximity is excluded in the global coordination step for emotion expression.

When the robot actually performs the generated deceptive action a_d, it must address possible conflict of a robot's actuators. Many bodily cues use the same joint, and it leads to the conflict if some of those cues are intended to be performed at the same time. To avoid this conflict, an integration step prioritizes among bodily cues that possibly use the same joints/motors. Formally, time-variation t is added to non-verbal action cues such as $< g_d^t, f_d^t, p_d^t >$. Time variable t represents the time to start the current action cue.

Therefore, if the potential conflict in actuator usage exits, t in each cue should be controlled. Proximity changes possibly involve the same joints/motors, as do some body gestures. Therefore, when a robot performs the action a, we prioritize proximity. Facial expression is obviously performed independently from the other cues, gesture and proximity, as there is no conflict in actuator usage. Therefore, a robot maintains facial expression during the performance of the proximity and gesture cues. Summarizing, as shown in Algorithm 1, a robot performs proximity cues first and then, if needed, produces the gesture cue while maintaining the facial expression

cue during the entire action.

In sum, in the integration module, a robot first generates the set of all combinations of possible deceptive bodily, facial, and proximity cues and filters out the contradictory actions to get the deceptive action set. Then, a robot determines whether any of these action combinations include conflict by observing the overlapping use of body parts and prioritizes the proximity cue to avoid those conflicts. Finally, the robot can produce the set of deceptive actions such as $A_d = \{a_{d1}, a_{d2}, \ldots, a_{dn}\}$ needed for the task at hand.

4.3 Deceptive Action Selection Mechanism

The deceptive action generation model in the previous section for the *method* part (Figure 12). In this section, as a next step, the computational models for the *motive* and *opportunity* parts are discussed. It describes how a computational model enables a robot to choose a beneficial behavior from either the true or deceptive action set based on other-oriented deception.

Motives are the reasons why a robot should perform deception in a certain situation. Opportunity indicates the possibility of successful performance of deception to benefit the mark (the deceived one). In other words, through the specific computational model, a robot should be able to determine if the current moment is the right time to perform deceptive or true actions. For this process, it is required for a robot to predict whether its potential deceptive behavior can have the motive to help human partners in certain situations. In addition, a robot needs to calculate which behaviors can maximize the benefits when applying various true and deceptive behaviors. More specifically, since other-oriented deception focuses on the benefits to the mark, measuring and calculating these benefits received by the mark are key factors of the motives/opportunities model. However, since a human's reaction is difficult to predict, modeling the mark's payoffs is challenging.

The approach to this modeling can involve a cognitive architecture. Previous research has proposed computational models for a robot to determine a human's cognitive stance inspired by a human's cognitive architecture. For example, several cognitive architectures (e.g., SOAR [83] and ACT-R [7]) were used to model human behaviors and applied to robotic systems. These cognitive robots can generate intelligent behavior based on a processing architecture that enables a robot to learn and reason about behaviors in response to complex goals [163, 41]. However, these cognitive architectures do not currently include a human's reasoning mechanism of deception.

Rather than a cognitive architecture model, using a psychological approach, human deception has previously been modeled based on interdependence theory and game theory [176]. Briefly, this model enables a robot to determine when it should trigger deception by predicting the mark's behaviors. The algorithm mainly consists of two parts: *when* to deceive and *how* to deceive. First, a robot determines whether deception can truly affect its human partner's status. Judging from the degree of this value, deception can be warranted or not in different situations. After deciding to engage in deception, a robot requires algorithms for performing deception. A game-theoretic approach is used for this step. However, this model has limitations since its calculation is based on payoffs that are currently predefined by the model. To overcome this issue, a new action-selection model is presented for a robot to learn and reason about the payoffs adaptively.

Instead of predicting the mark's benefits directly from a predefined model, it is more realistic that the robot should learn and reason about the mark's benefits in different situations and use this knowledge for its future decision process. This is more reasonable since a human's payoffs when he/she is deceived are somewhat unpredictable and are dependent on the situation. In sum, it is essential to develop a computational model so that a robot's other-oriented deception capabilities is able

to adapt flexibly based on its ongoing and prior experiences.

It is necessary to build a computational model that enables a robot to choose a correct behavior from the true or deception action set in each situation based on motivations of other-oriented deception. More specifically, this computational model should be able to adapt a robot's action selection mechanism since modeling true/deceptive behavior selection varies significantly by domains and users. In other words, a learning-based computational model for a robot to adapt the situation-action selection mechanism is required in this robot system.

The learning method can be classified in two ways, which are eager-learning methods and lazy-learning methods [30]. Eager-learning methods find the generalization during training (e.g., reinforcement learning techniques), and therefore, they require more training time to converge to the optimal solutions. However, this long training time is sometimes impractical in many HRI situations. For example, with the robot rescuers in the SAR situation, we cannot perform the training interactions thousands of times to achieve convergence.

Compared to eager-learning methods, a lazy-learning method is performed at the instance-query time. In the initial stage, the system can use preloaded or previously acquired cases, and it can find and adapt the action solutions from the casebase without a training session. By observing how the situation changes, the cases can be adapted and added to the casebase through experience. Therefore, even though it may include some initial generality issues, it is more feasible when applied to my computational model and domains. For this reason, I build the computational model based on the case-based reasoning (CBR) mechanism, one of the well-known lazy-learning mechanisms [4].

CBR is one of the lazy-learning techniques that allow finding and adapting a previous solution by reusing/maintaining previous experience [4, 82]. CBR is a method to solve new problems based on the solutions of similar previous problems. For this

process, the instances of a situation-solution pair are stored in the memory and referred to as a "case." By comparing the current situation with previous cases in the memory, the system should retrieve the nearest case and adapt the selected solution from the case according to the current situation. Finally, by observing the result of the case-selection, new cases should be retained and updated for future use.

According to Kolodner [82], the CBR cycle can be also illustrated in terms of four process stages. Here, four stages are defined as:

1. Case retrieval: when the problem occurs, the best matching case is searched to retrieve an approximate solution.

2. Case adaption: the approximate solution from the case retrieval is adapted to the new problem.

3. Solution evaluation: either before or after the solution is applied, the adapted solution can be evaluated.

4. Casebase updating: based on the verification of the solution, the case should be updated or the new case may be added to the casebase.

CBR has already been applied in several different robotic systems successfully. For example, it has been used for learning parameterization for autonomous navigation tasks [86]. To reuse prior robot missions and redesign/repair robot missions better, CBR has also been successfully applied [105]. More recently, another study built a CBR model to maintain and learn affective robotic attitudes [104].

In this model, a robot should be able to select an appropriate action from the set of true and deceptive actions in a given situation. Therefore, the model should store the information of situation-action pairs that increase the human's benefits. Effective action selection is one of the obvious and efficient robotic domains that use the CBR techniques successfully. In previous work, retrieving and reusing previous

game plays for robot soccer utilized CBR [129, 128]. To increase a robot's learned action responses in HRI contexts, the CBR method has been proposed and applied successfully. Similar to previous research, I developed a computational model for a robot to learn and select the most appropriate true/deceptive action via CBR [142].

In the rest of this section, details on how CBR is used in this action selection model will specifically be explained (in Section 4.3.1). After explaining computational details, a specific situation where other-oriented deception can be useful will be chosen, and this model will be explained using the exemplar scenario (in Section 4.3.2).

4.3.1 Deceptive Action Selection via CBR

4.3.1.1 Case C

For the CBR architecture, a previous experienced situation-action set should be stored in the memory. In addition, how much the action can benefit the human partner should be contained in the case. Therefore, a situation-action-benefits trio is required to define a case c in my CBR mechanism, and those cases should be retrieved and adapted for reuse during the CBR process. As a result, a case c consists of a situational state s, a corresponding action a, and the resulting benefits r, and the set of cases can be defined as follows:

$$C = \{< s, a, r > | s \in S, a \in A, r \in R\}$$

State S

First, the input situational state should express the current state of the mark during the interaction and also the environment. A robot should observe the current situation as an input and then should select and perform an appropriate true/deceptive action as an output. Therefore, state s is defined as the combination of features, which can represent the mark's internal $(f_{m1}, f_{m2}, \ldots, f_{mj})$ and environmental $(f_{e1}, f_{e2}, \ldots, f_{ek})$ conditions. Finally, the set of situational state for this CBR

mechanism can be defined as shown below:

$$S = \{< f_{m1}, f_{m2}, \ldots, f_{mj}, f_{e1}, f_{e2}, \ldots, f_{ek} > |$$
$$f_{mi} = \text{features to perceive the mark's internal conditions,}$$
$$f_{ei} = \text{features to perceive the mark's environmental conditions}\}.$$

Please note that features can be valued as *'don't care'* if the system determines that they do not play a key role in state discrimination. More details about the use of *'don't care'* features will be shown later (in *'Similarity Score"* subsection).

Action A

In section 4.2, a novel algorithm to generate a robot's action set, which includes the true actions and alternative deceptive actions, has been discussed. Briefly, a robot can find and generate the action set based on the general/emotional action primitives. A robot action is defined as the combination of different action cues; $a =< g, f, p >$. Here, g is the bodily gesture cue, f is the facial expression cue, and p is the proximity cue. When the robot's true action a_t is determined, a robot can generate the deceptive actions based on the characteristics of each primitive. This deceptive action generation is possible according to two mechanisms, which are deception by commission and deception by omission. Finally, from the default n true actions and alternative m deceptive actions, the total set of actions A can be defined such as $A \in \{a_{t1}, a_{t2}, \ldots, a_{tn}, a_{d1}, a_{d2}, \ldots, a_{dm}\}$.

Benefit R

The main goal of this model is to find the robot's action that can maximize the benefit of the deceived human partner (the mark) in each situation. Therefore, how much benefit the mark can get from a robot's true/deception action is an essential element for selecting the case. Here, the measure of those human benefits is defined

as R. To determine the benefits for the deceived human partner, a robot should observe how the state is changed from the perceived situational state when an action is performed. In general, we can anticipate the result of a robot action such as whether the situation is getting better, worse, or just staying the same.

The benefit R is defined using numeric measures. Benefits can vary based on the situation changes. The situation can be getting better, worse, or maintained, therefore the level of situation changes can be described by the set of integers. From this aspect, if the situation is getting worse, the benefit will be measured with negative integers. Similarly, positive integers can represent the degree of improved situation. Obviously, 0 will illustrate that the situation is just staying the same. As a result, the set of benefits R can be defined as:

$$R = \{r | r \in \mathbb{Z} \ \wedge \ R_w \leq r \leq R_b\}$$

Here, $\mathbb{Z}$ is the set of integers. R_w is the minimum integer for the negative benefits and R_b is the maximum integer for the positive benefits.

Example: Case definition in problem-solving task situation

Example Situation: Using other-oriented deception with a robotic educational assistant to help students complete problem-solving task

Situation Description: Let us assume that a student (the mark in this example) has to solve the sequence of problems in an educational setting. To help the student, a robot assistant placed next to the student can provide feedback on his/her performance, namely whether he or she is correctly or incorrectly answering questions.

Robot Deception: Research has suggested that students can be motivated by the deceptive reactions of teachers. Even if a student's performance is poor, by showing positive, deceptive, feedback teachers can motivate students.

Example: How to define case $c =< s, a, r >$ in this example

1. State S: The decision can be made based on the mark's current and previous performance, and its motivation can be a determining factor in detecting a student's internal status. In that sense, the mark's current emotional status is an essential feature.

Example State definition: $s =< f_{emotion}, f_{shortterm}, f_{longterm} >$ where

$f_{emotion} = \{positive, negative, neutral\}$; the mark's emotional state

$f_{shortterm} = \{x | x \in \{correct, incorrect\}\}$; short-term performance (correctness in the previous question)

$$x = \begin{cases} correct, & \text{if the mark solves the previous question correctly} \\ incorrect, & \text{otherwise} \end{cases} \quad (5)$$

$f_{longterm} = \{x | x \in \mathbb{R} \ \wedge \ 0 \leq x \leq 100\}$; long-term performance (percentage of the correctness from the beginning of the first question to the current question)

2. Action set A: The robot's action is feedback upon the student's current performance. If the student's solution is correct, then the robot makes a positive (happy) gesture. If student's answer is incorrect, however, then it makes a negative (sad) one. Between those actions, either true or deceptive actions are selected for each state.

Example Action Set: $A = \{< g_{positive}, f_{positive}, \textit{don't care} >, < g_{negative}, f_{negative}, \textit{don't care} >, < g_{neutral}, f_{neutral}, \textit{don't care} >\}$

3. Benefit R: The mark's benefits can be determined by observing whether

its performance improves, worsens, or does not change. By comparing the performance on the previous and the current questions, the response can be determined as shown below.

Example Benefit Definition: $R = \{r | r \in \mathbb{Z} \ \wedge \ -1 \leq r \leq 1\}$ where

$r = -1$: answer correctly in previous problem but incorrectly in current problem (worsen)

$r = 0$: answer correctly in both previous and current problem OR answer incorrectly in both previous and current problems (maintained)

$r = 1$: answer incorrectly in previous problem and correctly in current problem (improved)

4.3.1.2 Case-based Reasoning Process

Overall, the deceptive action selection model is based on on Kolodner's CBR cycles work [82]. The case-based reasoning process consists of the following steps: case retrieval, case adaption/reuse, and evaluation/updating.

- Case retrieval: when the problem occurs, the best matching case is searched to retrieve an approximate solution.
- Case adaption and Reuse: the approximate solution from the case retrieval is adapted to the new problem and the adapted solution is applied.
- Solution Evaluation and Casebase updating: after the solution is applied, the adapted solution is evaluated and based on the evaluation result, the case should be updated or the new case may be added to the casebase.

Figure 22 illustrates the overview of this section selection model using CBR. Briefly, the robot should first perceive the current status and determine the most similar case from the casebase (step 1. case retrieval). After adapting the selected case's action, the robot performs the adapted action (step 2. case adaptation and reuse). Finally,

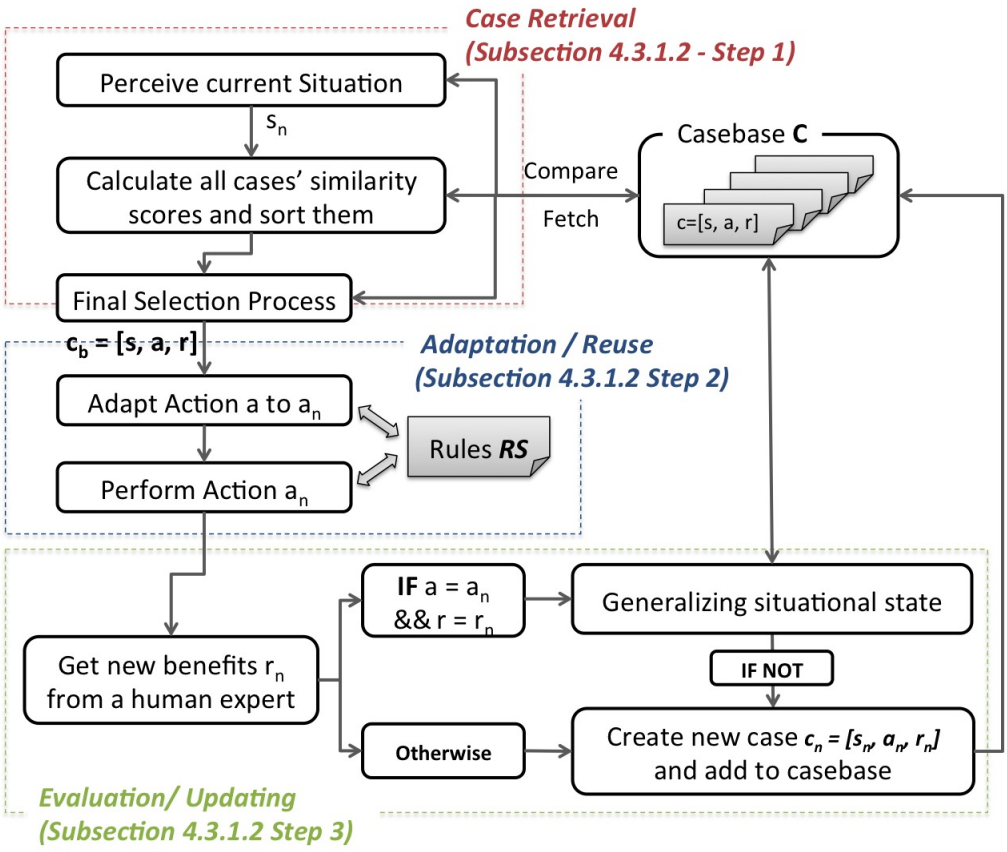


Figure 22: Overview of Action Selection Model using CBR

by calculating benefits of the mark, the casebase should be updated for future uses (step 3. solution evaluation and casebase updating). Details of each step will be illustrated in the following sub-sections.

Step 1. Case Retrieval

In the case retrieval step, the system should find which case(s) has the most similar situational state to the current perceived state. The case retrieval algorithm scores how similar the current situational state is to other cases in the casebase. By calculating this similarity score, the algorithm can select the best-matched case or cases in this step.

Similarity metric

To determine the initial matched cases, a similarity score is calculated that measures how much the state in the case is similar to the current situation. For this, a syntactic similarity assessment, which provides a global similarity metric based on surface match, is used. For computing this score, a Manhattan distance calculation is used, similar to [104], by calculating the $L-1$ norm between the two feature vectors of the situational states. Since not all features in the state are equally important, each feature is weighted to indicate its relative importance. $\Phi(s_p, s_c)$ is defined as the similarity metric between states s_p and s_c as shown in equation 6 when the state vectors are illustrated as $s_p =< f_{1,p}, f_{2,p}, \ldots, f_{n,p} >$ and $s_c =< f_{1,c}, f_{2,c}, \ldots, f_{n,c} >$.

$$\Phi(s_p, s_c) = \frac{\sum_{i \in S} k_i \cdot w_i}{\sum_{i \in S} w_i} \tag{6}$$

Here, k_i is a similarity score for each feature f_i and w_i is a weight for each feature f_i. The feature valued *'don't care'* in s_p indicates that it can be matched with any feature value in state s_c. Therefore, the similarity score for this feature can be the maximum value. Weight w_i should be empirically set by a human expert's domain knowledge. Similarity score k_i is defined as $k_i = \{x | x \in \mathbb{R} \ \wedge \ 0 \leq x \leq 1\}$ and it can be calculated differently based on the characteristics of each feature f_i as shown in

Algorithm 2 Calculating similarity score k_i for feature f_i

1: **if** feature f_i is *'don't care'* **then**
2: $\quad k_i = 1$
3: **else if** feature f_i is a numeric datum **then**
4: $\quad k_i = 1 - \frac{|f_{i,c} - f_{i,p}|}{v_i}$ where v_i is the range for feature f_i
5: **else if** feature f_i is a categorical datum **&&**
Cardinality of set $F_i \leq 2$ where $f_i \in F_i$ **then**
6:

$$k_i = \begin{cases} 1 & \text{if } f_{i,p} = f_{i,c} \\ 0 & \text{otherwise} \end{cases}$$

7: **else if** feature f_i is a categorical datum **&&**
Cardinality of set $F_i \geq 3$ where $f_i \in F_i$ **then**
8: $\quad k_i = M_i(f_{i,p}, f_{i,c})$ where M_i is the f_i-specified similarity table
9: **end if**

Algorithm 2.

As described in Algorithm 2, when feature f_i is a categorical datum, categorical matching can be used. However, if there are more than two categories, there may be shades of similarity; one category might be more like the one of the other categories than another. To specify those shades of similarity the feature-specified lookup table M_i is defined according to the characteristics of each feature, and similarity scores for feature f_i can be determined by the corresponding entry in this lookup table M_i.

After calculating $\Phi(s_p, s_c)$ for each case in the casebase, the cases are sorted from the highest scored case to the lowest one, and the top n cases are selected as the initial matched cases and transferred to the case adaptation step.

Example: Calculating similarity score

Let us assume $s_{sp} =< negative, incorrect, 70 >$ and $s_{sc} =< positive, incorrect, 50 >$ where $s = < f_{emotion}, f_{shortterm}, f_{longterm} >$ as described in "Example: Case definition in problem-solving task situation." Similarity scores for each feature can be calculated according to Algorithm 2 as shown below.

FOR each feature f_i in $i = emotion, shorterm, longterm$

IF $i == emotion$ *//Cardinality of set* $F_{emotion} \geq 3$

Find k_i from similarity lookup table $M_{emotion}(f_{emotion,sp}, f_{emotion,sc})$

$\rightarrow k_{emotion} = M_{emotion}(negative, positive) = 0$

ELSE IF $i == shortterm$ *// Cardinality of set* $F_{shortterm} \leq 2$

IF $f_{i,sp} == f_{i,sc}$ **THEN** $k_i = 1$

ELSE $k_i = 0$

$\rightarrow k_{shorterm} = 1$

ELSE IF $i == longterm$ *//Feature* $f_{longterm}$ *is a numeric datum*

$k_i = 1 - \frac{|f_{i,sc} - f_{i,xp}|}{v_i}$ where v_i is the range for feature f_i

$\rightarrow k_{longterm} = 1 - \frac{|50-70|}{100} = 0.8$

Please note that similarity lookup table for emotion $M_{emotion}$ is defined as follows in this example.

	positive	neutral	negative
positive	1	0.5	0
neutral	0.5	1	0.5
negative	0	0.5	1

Finally, based on the results of similarity scores for each feature, the similarity score between two states can be calculated as follows. Please note that it is assumed that all features have equal weights such as $w = \{1, 1, 1\}$.

$$\textbf{Similarity score } \Phi(s_{sp}, s_{sc}) = \frac{\sum_{i \in S} k_i \cdot w_i}{\sum_{i \in S} w_i} = \frac{(0+1+0.8)}{1+1+1} = \mathbf{0.6}$$

Step 2. Case adaption and reuse

Once the initial matched cases are selected, the final selection process is required to find the best matched case $c_b = [s, a, r]$. Then, the corresponding action in this best matched case should be adapted and passed to the robot to perform the robot behavior. Action a is one of the actions among the true and deceptive action sets. However, the cases learned from the experts or previous experience cannot guarantee the best solution in the CBR model. In other words, if the selected action in the existing case causes the situation to become worse, it is necessary to adapt and select a new action to find a better solution.

Even though a robot finds the best matched situation, there is no reason to reuse the action if the benefits of the action are bad. Therefore, as a final selection step, a robot should assess the previous benefits r and determine whether the previous action a will be directly performed or need to be adapted.Therefore, the final action selection mechanism as shown in Figure 23.

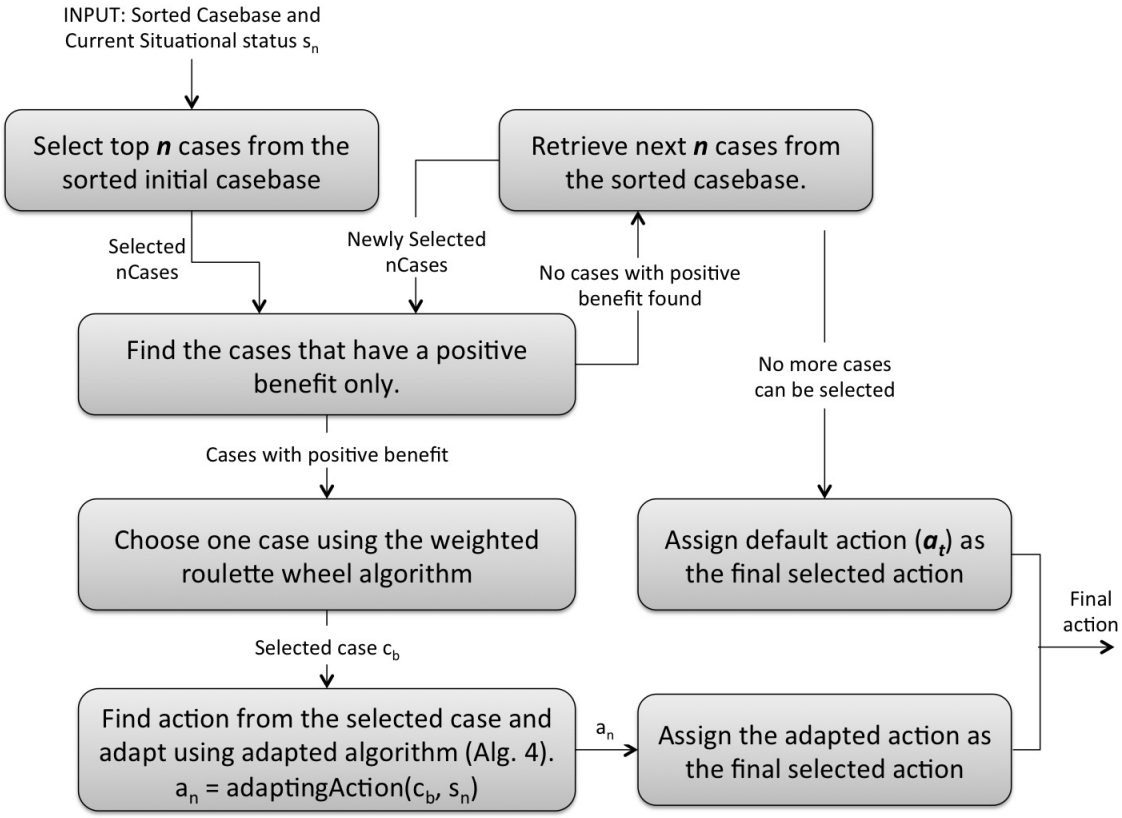


Figure 23: Overview of final action selection in case adaptation

More specifically, this final action selection process can be illustrated using the pseudocode shown in Algorithm 3. Again, after the best-case is determined ($c_b = [s, a, r]$), it may need to be adapted to the current situation. If the current state is exactly matched to the situational state in the best-case, we can directly use action a in the case without adaptation. However, if those two states are slightly different but the case is selected since it has the most similar situational state, then the action may be adapted before application.

Algorithm 3 Final Action Selection Process

Input: Sorted Casebase C, Current Situation s_n
Output: Final adapted solution action a_n

```
1: startIndex = 0; // starting index of casebase when picking top n cases
2: LOOP: // start the loop
3: a_n = a_t
4: nCase = {} // a variable to store candidate cases
5: for i in startIndex to startIndex+n // find candidate cases do
6:    if c[i].r > 0 // if the benefit of the case is positive then
7:       nCases.add(c[i]) // that case is added to the set of potential best cases (candidate cases)
8:    end if
9: end for
10: if |nCases| > 0 // if potential best cases exist then
11:    Find the best case from nCases using the weighted roulette wheel algorithm
12:    c_b = selected best-case // randomly select one best-case from nCase
13:    a_n = adaptingAction(c_b, s_n) // adapting the action (Algo. 4)
14:    return a_n // find final adapted action
15: else
16:    startIndex += n // to look at the next top n cases from the casabase
17:    goto LOOP // repeat the loop
18: end if
19: a_n = a_t // if best-case can not be founded to the end of the casebase,
20: return a_n // set the default action a_t as the final adapted solution
```

As shown in Algorithm 3, the selected action a should be adapted if necessary. According to Kolodner [82], adaptation involves two major steps: 1) figuring out what needs to be adapted and 2) doing the adaptation. Similar to this approach, the algorithm first observes whether there are differences between best-case's state s and

Table 10: Data structure for the adaptation rule

Field	**Description**
FeatureIndex	Indicates which feature should be observed for the current adaptation rule
Origin	The origin of the adaptation rule (reference)
Activity	Indicates if the rule is currently active
Description	Short, concise description of the rule
AdaptationCondition	Condition for determining whether the adaptation rule should be applied to the current action
AdaptationRule	Formal expression defining the way to adapt the current action

current state s_n, and if two states are different each other, an appropriate adaptation should be applied. Differences between states can be discriminated by figuring out which features specifically have different values (step 1. Figuring out what needs to be adapted). When the different features are determined, the adaptation rules related to the discriminated features are applied to the action (step 2. Doing the adaptation).

Then, how can the adaptation rules be defined? As stated in Kolodner's mechanism [82], adaptation can be varied for any particular domains or tasks, and a set of adaptation strategies or heuristics can be used for a CBR working system. Therefore, the adaptation process can also be designed by the predefined adaptation rules.

For this purpose, the data structures of the rules that encode the adaptation procedures (from experts or literature) are defined. Here, the adaptation rules give an instruction for a robot to adapt the selected action to the current situation. The structure of the rules is defined as shown in Table 10. Again, the rules for adaptation will be predefined according to the characteristics of the context.

The adaptation process using the rules is summarized in Algorithm 4. As illustrated in this algorithm, when the situational state in the selected best-case is exactly same as the current situation, there is no reason to adapt the solution, and therefore,

Algorithm 4 Adaptation Process: *AdaptingAction*(c_b, s_n)

Input: best-case $c_b = [s, a, r]$ where $s =< f_{1,s}, f_{2,s}, \ldots, f_{k,s} >$
Current state $s_n = < f_{1,sn}, f_{2,sn}, \ldots, f_{k,sn} >$
Set of rules $RS = \{rs_1, rs_2, rs_3, \ldots, rs_n\}$
Output: Adapted action a_n

```
1: // determine whether the current state is same as the case's state
2: if s = s_n then
3:    // if two states are same,
4:    a_n = a // set the same action a as an adaptation action
5: else
6:    // if two states are different, start adaptation
7:    D = {} // variable to store the set of feature indices
8:    // Step1. Figuring out what needs to be adapted
9:    for all features in s do
10:      if f_{i,s} != f_{i,sn} then
11:         // if the value of feature f_i in case's state s is different from current state
            s_n, store feature index i into the set D
12:         D ← D ∪ i
13:      end if
14:   end for
15:   // Step2. Doing the adaptation
16:   for all feature index i in D do
17:      // among all rules in RS
18:      for all rules rs_j in RS do
19:         if rs_j.featureIndex = i AND rs_j.adaptationCondition = TRUE then
20:            rs_j.adaptationCondition = true
21:            // find the rule rs_j that contains the feature index i
22:            // if rs_j's adaptation condition is satisfied, adapt by the rule
23:            a_n ← Perform rs_j.adaptationRule
24:         end if
25:      end for
26:   end for
27: end if
28: return a_n
```

the action a in the case is directly used as a solution. However, when the two situations are different, it is necessary to discriminate what are the differences between two situational states s and s_n, then based on these differences, appropriate adaptation should be applied. The situation state s is defined by several features such as $< f_1, f_2, \ldots, f_k >$. Therefore, the system should first compare all feature values between the case state and the current state. If the two values are different, those feature indices are stored in set variable D. After the features that have different values are discriminated in D, those values are compared to the element in the set of rules RS. In other words, if the element in D is matched with the *FeatureIndex* in certain $rs \in RS$, this rule rs's *AdaptationRule* applies to the selected best-case action a based on *AdaptationCondition* and finally the adapted action a_n will be generated.

Finally, from the retrieval and adaptation steps, a robot can determine an appropriate adapted action a_n. Then, this adapted action a_n should be performed by the robot. This step is known as the case application step. When the adapted case is applied (action a_n is determined and performed), the changes of states should be observed for the next evaluation/updating steps.

Example: Adaptation Rule and Process

Example Adaptation Rule: If the longterm performance is extremely low (bad), no deception is necessary.

<rule> $rs_{extreme_low_performance}$

<feature Index> **longterm** <feature Index>

<origin> **Pigmalion Effects** <origin>

<active> **true** <active>

<description> Perform the true action if student's performance is continuously and extremely low. <description>

<adaptation Condition> $f_{longterm} \leq 20$ <adaptation Condition>

<adaptation Rule> $a_n = a_t$ <adaptation Rule>

If exemplar inputs are defined as selected best case $c_b = [< negative, incorrect, 70 >, a_{d1}, 0]$ and current state $s_n = < positive, incorrect, 20 >$, the exemplar adaptation rule can be applied using Algorithm 4 as shown below.

Adaptation Algorithm	Exemple
IF $s == s_n$ $a_n = a$	<negative, incorrect, 70> != <positive, incorrect, 20>
ELSE D = {} **FOR** all features in s **IF** $f_{i,s}$!= $f_{i,sn}$ **THEN** D ← D U { *i* }	D = {}; // values of feature $f_{emotion}$ and $f_{longterm}$ are different D ← {*emotion*, *longterm*}
FOR all feature index *i* in **FOR** all rules rs_j in RS **IF** (rs_j.featureIndex == *i* && rs_j.adaptationCondition == True) a_n ← Perform rs_j.adaptationRule;	$rs_{extreme_low_performance}$ contains indices *longterm* && adaptation Condition == True ($f_{longterm} \leq 20$) Perform the adaptation Rule ➔ $a_n = a_t$

Step 3. Solution Evaluation and Casebase updating

Cases are initially created by experts or experienced users. Even then, the pair of situational state and action in each case may not always be the best solution. During the evaluation/updating phase, the cases should be revised and updated to the most effective solution. In other words, it should be observed whether a deceived human partner truly receives benefit from a robot's behavior, and the casebase should be updated with the case that can generate the highest payoff for humans.

After performing an action (case application), the benefits should be identified to evaluate the action post facto. It can be determined by a human expert afterwards or if possible assessed autonomously by a robot. In the current system, a human expert will rate the benefits since it is a more practical and accurate way in the real-world situation. When the human expert determines the benefits, they should be asked to observe the changes of the human's state after performing a robot's adapted action and rate the new benefit r_n within the predefined range of benefits R.

Updating casebase

Based on the new benefit r_n, which is provided by a human expert, the CBR system should update the cases if necessary. According to Kolodner [82], the casebase

can be updated by 1) generalizing the cases in the casebase, or 2) storing the new case. Similar to this approach, when the current state s_n, the adapted action a_n, and the new benefit r_n are perceived, they can be used to generalize existing cases in the casebase. Otherwise, it will be stored as the new case for future reference.

First, the system should determine whether generalization is possible with s_n, a_n, and r_n. In this model, states can be generalized when the cases have the same action as a_n and benefit as r_n. Since states are represented as different features, the system can generalize the state by finding/merging the features that may not play key roles during the calculation of similarity scores. In other words, for the generalization process, those unnecessary features can be minimized to a *'don't care'* value.

How to determine *'don't care'* features and generalize the casebase

This feature reduction can be solved inspired by the minimization algorithm of the algebraic variables [181]. In algebra, variables can be represented via algebraic definition and operations. For example, in the boolean algebra, values can be represented by variables with NOT operation. If X and Y are boolean variables and values of these variables are $X = 1$ and $Y = 0$, then these values can be represented as X and $\bar{Y}$. In addition, when one term is mixed with multiple variables, it is presented as the product form. Again, if the multi-variate term contains $X = 1$ and $Y = 0$, it can be represented as $X\bar{Y}$ using algebraic representation. Such product representation is also called *minterm*.

If several minterms are included in one equation, then the equation can be represented using the form of the *sum of products (SOP)*. When multiple minterms appear in the SOP expression, some variables are not necessary to achieve the same answer; termed *'don't care.'* features. For example, let us assume that two terms are $XY\bar{Z}$ and XYZ. Using the SOP, it can be expressed $XY\bar{Z} + XYZ$. This equation's answer is only depends on XY, and Z cannot influence the result ($XY\bar{Z} + XYZ =$

$XY(\bar{Z}+Z) = XY$ since $\bar{Z}+Z$ is always true). Therefore, variable Z is determined as *'don't care'*, and the equation can be minimized as XY. In other words, from the midterm XY, the *'don't care'* variable Z can be also detected.

Similar to this approach, *'don't care'* features can be discriminated in the model. All features in the state can be represented using algebraic variables and SOP forms, and once the SOP equation is determined, by minimizing minterms, the *'don't care'* features can be determined. Finally, by minimizing those *'don't care'* features, the casebase can be minimized and generalized.

When all the features are boolean variables, the *'don't care'* feature can easily be determined by a Karnaugh mapping [182]. However, since this method is based on tabular form, it is difficult to make efficient for use in computer algorithm. Instead, the Quine-McCluskey (Q-M) algorithm can be used [97]. Q-M algorithm is a deterministic method to find the minimal form of a Boolean expression. Briefly, the algorithm works in two steps. First, all prime implicants of the boolean expression is determined. Then,the essential prime implicants are discriminated using prime implicant chart. The details of pseudo-codes and implementation appear in McCluskey's paper [97]. The following example is shown the minimization/generalization process for three states using this Q-M algorithm.

Example: States Generalization 1

Let us generalize three states $s_1 =< 1,1,1 >$, $s_2 =< 1,1,0 >$, and $s_3 =< 1,0,1 >$ where all features are boolean variables.

1. Represent three states as minterms

   $\rightarrow s_1 = f_1 f_2 f_3$, $s_2 = f_1 f_2 \bar{f}_3$, $s_3 = f_1 \bar{f}_2 f_3$

2. Find SOP form

   $\rightarrow f_1 f_2 f_3 + f_1 f_2 \bar{f}_3 + f_1 \bar{f}_2 f_3$

3. Find minimized SOP form using Q-M algorithm

$\rightarrow f_1 f_2 f_3 + f_1 f_2 \bar{f}_3 + f_1 \bar{f}_2 f_3 = f_1 f_2 + f_1 \bar{f}_2 f_3$

4. Discriminate *'don't care'* features from the minimized minterms

   $\rightarrow f_1 f_2 =$ <1, 1, *'don't care'*>

   $\rightarrow f_1 \bar{f}_2 f_3 + f_1 \bar{f}_2 f_3 =$ <1, 0, 1>

Finally, three states s_1, s_2, s_3 can be minimized to two states <1, 1, *'don't care'*> and <1, 0, 1>.

Multi-valued features can be handled by the same mechanism via the extended Q-M algorithm [101]. Again, the extended Q-M algorithm allows to identify the minimization form for multi-valued algebraic functions. The pseudo-codes and implementation details appear in Mishchenko's paper [101], and the exemplar use of this minimization/generalization process can be shown as follows.

Example: States Generalization 2

Let us generalize four states $s_1 =< 1, a, x >$, $s_2 =< 1, b, y >$, $s_3 =< 1, a, y >$, and $x_4 =< 1, c, y >$ where features are defined as $f_1 \in \{1, 2, 3\}$, $f_2 \in \{a, b, c\}$, $f_3 \in \{x, y\}$.

1. Represent three states as minterms

   $\rightarrow s_1 = f_1^1 f_2^a f_3^x$, $s_2 = f_1^1 f_2^b f_3^y$, $s_3 = f_1^1 f_2^a f_3^y$, $s_4 = f_1^1 f_2^c f_3^y$

2. Find SOP form

   $\rightarrow f_1^1 f_2^a f_3^x + f_1^1 f_2^b f_3^y + f_1^1 f_2^a f_3^y + f_1^1 f_2^c f_3^y$

3. Find minimized SOP form using extended Q-M algorithm

   $\rightarrow f_1^1 f_2^a f_3^x + f_1^1 f_2^b f_3^y + f_1^1 f_2^a f_3^y + f_1^1 f_2^c f_3^y = f_1^1 f_2^a f_3^x + f_1^1 f_3^y$

4. Discriminate *'don't care'* features from the minimized minterms

   $\rightarrow f_1^1 f_2^a f_3^x =$ <1, a, x >

   $\rightarrow f_1^1 f_3^y =$ <1, *'don't care'*, y >

Finally, four states s_1, s_2, s_3, s_4 can be minimized to two states <1, a, x > and <1, *'don't care'*, y >.

Algorithm 5 Casebase Updating Strategy

Input: Current State s_n
Adapted Action a_n
Output: Updated Casebase

```
1: // Determine the new benefit r_n from a human expert
2: // Step 1. Generalizing the cases
3: candidateCases = {} {cases that are potentially generalizable}
4: for each case c in Casebase C do
5:    if c.r = r_n && c.a = a_n then
6:       // if the case has the same action and benefits as a_n and r_n, it is potentially
         generalizable
7:       candidateCases.add(c) // add to the candidateCases
8:    end if
9: end for
10: // Generalize states via minimization algorithm
11: allMinterms = ExtendeQ-M (canonical forms of candidateCases' states, canonical
    form of s_n)
12: for each minterm_i in allMinterms do
13:    s_generalized = extract from minterm_i by adding 'don't care' terms
14:    Remove all cases in candidateCases
15:    Add generalized case [s_generalized, a_n, r_n]
16: end for
17: // Step 2. Storing the new case
18: if no cases are generalized then
19:    Create new case c_n = [s_n, a_n, r_n] // new case created
20:    Add c_n to the case base // updated
21: end if
```

Algorithm 5 describes the entire process of this updating strategy in pseudocode. Again, for the generalization process, the system first selects all cases that have the same action as the adapted action a_n and the same benefit as the new benefit r_n. The states from the selected cases and the current state s_n are then generalized by reducing the features. This feature reduction is performed using the minimization algorithm [97, 101]. After the reducible features are determined, those features are valued as *'don't care'* and the cases are generalized. If none of the cases are merged

and generalized in the previous step, a new case $c_n = [s_n, a_n, r_n]$ is created and stored in the casebase for future use. By using this updating strategy, the casebase can maintain the best situation-action pairs that can provide the largest benefit to the deceived humans.

Example: Updating Casebase

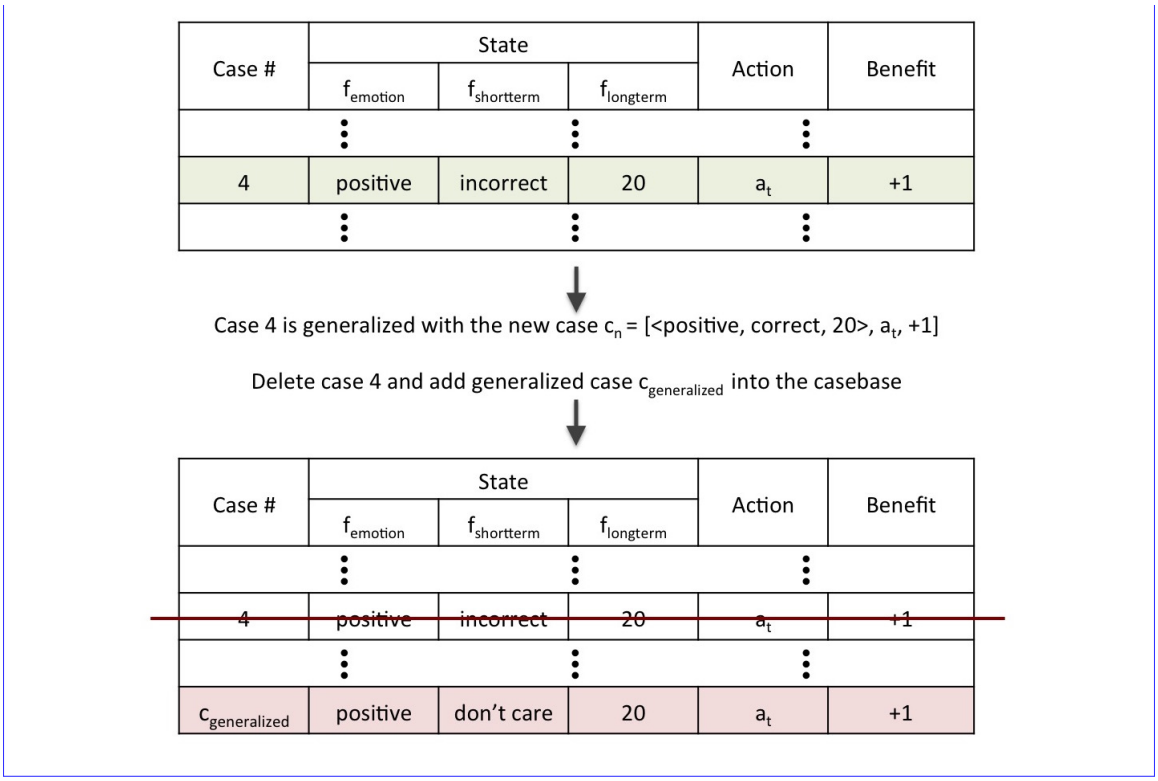
After performing the adapted action, the new case will be added to the casebase or generalized with the current case(s) in the casebase. Let us assume that the final case is $c_n = [< positive, incorrect, 20 >, a_t, +1]$. The generalization/updating strategy is performed as shown below.

Algorithm 4	Example
Determine new benefit r_n from an expert	Let's assume that the new benefit r_n = +1

/* Step 1. Generalizing the cases */
candidateCases = {};
FOR each case c in Casebase C
  IF c.r == r_n && c.a == a_n
  candidateCases.add(c);
allMinterms = ExtendedQ-M ();
FOR each minterm$_i$ in allMinterms
  $s_{generalizeed}$ = extract from minterm$_i$ by adding '*don't care*' terms
  Remove all cases in candidateCases
  Add generalized case [$s_{generalizeed}$, a_n, r_n]
	currentCase = {[<*positive*, *incorrect*, *20*>, a_t , +1]}
	
	candidateCases = {[<*positive*, *incorrect*, *20*>, a_t , +1]} , [<positive, correct, 20>, a_t , +1]}
	
	Find allMinterms = $f_{emotion}^{positive} f_{shortterm}^{incorrect} f_{longterm}^{20} + f_{emotion}^{positive} f_{shortterm}^{correct} f_{longterm}^{20}$
	
	➔ Minterm = $\{f_{emotion}^{positive} f_{longterm}^{20}\}$
	➔ Generalized case [<*positive*, *don't care*, *20*>, a_t , +1]

/* Step 2. Storing the new case */
IF no cases are generalized
  Create new case c_n = [s_n, a_n, r_n]
  Add c_n to the case base
	/* Create and Add generalized case */
	Add generalized case c_n = [<*positive*, *don't care*, *20*>, a_t , +1] to the casebase

Finally, by generalizing the cases, the final casebase can be updated as shown below.



4.3.2 Exemplar Scenario

The deceptive action-selection model presented in Section 4.3.1 can be summarized as shown in Figure 24. This computational model enables a robot to perceive the current situation and to choose and maintain the action among alternative true and deceptive actions. The goal of this subsection is to review this motive/opportunity model with a specific example. One significant domain where we can assume to use other-oriented deception is in the search and rescue situation. According to Lois' case study [89] and International Association of Fire Chiefs' (IAFC) manual [2], it is essential for human rescuers to manage a victim's emotions during the crisis situation, and for this purpose, other-oriented deceptions are sometimes used to calm victims fears. Under this circumstance, the search and rescue domain is selected as an example. There exist many different types of search and rescue contexts. More specifically, the scene of fire is chosen as the exemplar scenario in this subsection.

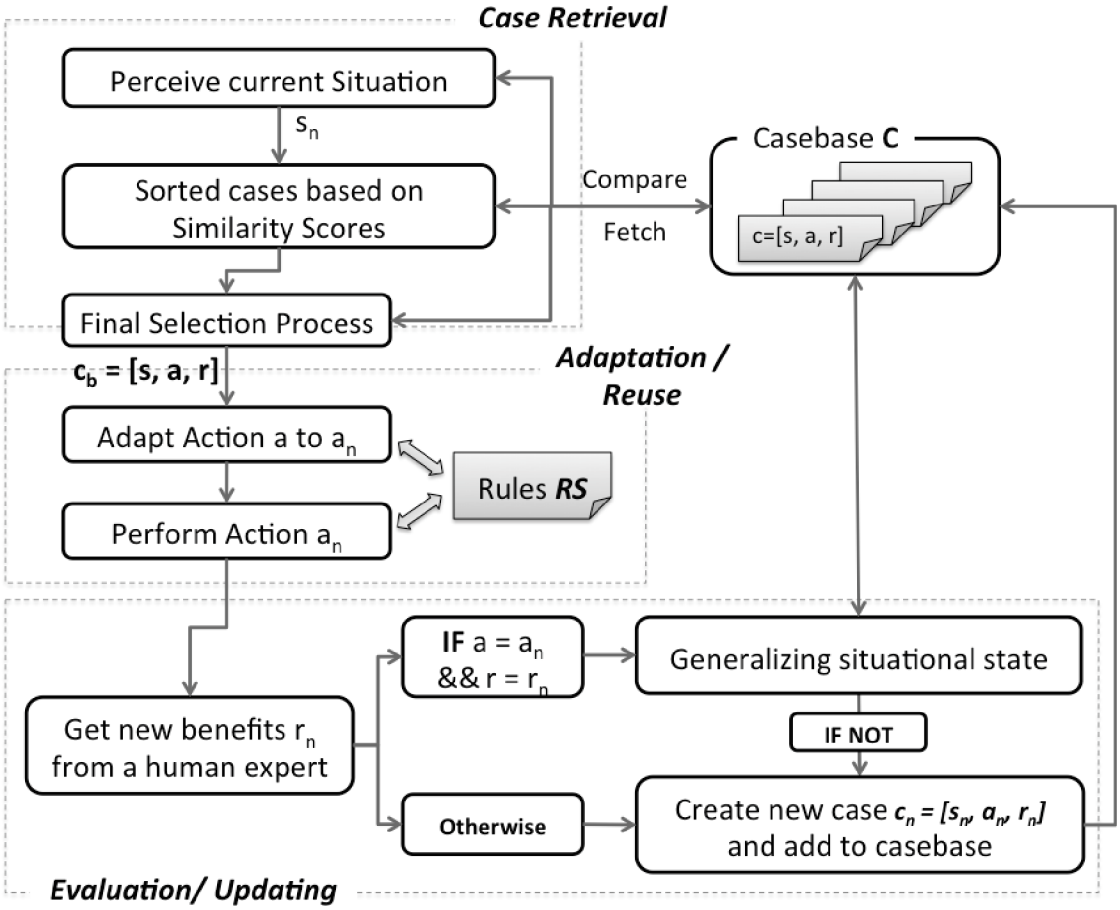


Figure 24: Computational architecture for the motives/opportunities model

Table 11: Exemplar Scenario: Data structure of state S

Feature	Description
$f_{respiration}$	$\{x \mid x = 1 \text{ if } 10 < r_{respiration} < 30, otherwise\ x = 0\}$: Normal or abnormal respiration : $r_{respiration}$ from respiratory sensor (breaths per minute, bpm)
f_{pulse}	$\{x \mid x = 1 \text{ if } 60 < r_{pulse} < 100, otherwise\ x = 0\}$: Normal or abnormal ranges of pulse : r_{pulse} from pulse rate sensor (pulse beats per minute, $pBPM$)
$f_{heartbeat}$	$\{x \mid x = 1 \text{ if } 60 < r_{heartbeat} < 100, otherwise\ x = 0\}$: Normal or abnormal range of heartbeat : $r_{heartbeat}$ from heart rate sensor (heart beats per minute, $hBPM$)
$f_{emotion}$	$\{x \mid x \in \{anger, disgust, fear, happiness, sadness, surprise\}\}$: Current emotional state from speech and pitch detection
$f_{temperature}$	$\{x \mid x = 1 \text{ if } 14 < r_{temperature} < 32, otherwise\ x = 0\}$: Normal or abnormal range of room temperature : $r_{temperature}$ from digital temperature sensor (Celsius, $^{\circ}C$)
f_{gas}	$\{x \mid x = 1 \text{ if } r_{gas} == false, otherwise\ x = 0\}$: Gas detected or not, r_{gas} from CO-gas detection sensor

Case C

In the opportunities/motives model, case is defined as $c = [s, a, r]$, where s is a situational state, a is an action, and r is the benefit. In this search and rescue example, the mark's state should represent/cover the human victim's physical/emotional conditions. As always, to ensure reasonable convergence times, the number of states should not be too large. Here, the mark's conditions are determined by four internal features and two external features, and so, state s can be defined as $s =< S_{physical}, S_{environmental} >=< f_{respiration}, f_{pulse}, f_{heartbeat}, f_{emotion}, f_{temperature}, f_{gas} >$ where $S_{physical} = \{< f_{respiration}, f_{pulse}, f_{heartbeat}, f_{emotion} > | f_{respiration} \in \{1, 0\}, f_{pulse} \in \{1, 0\}, f_{heartbeat} \in \{1, 0\}, f_{emotion} \in \{\text{anger, disgust, fear, happiness, sadness, surprise}\}\}$ and $S_{environmental} = \{< f_{temperature}, f_{gas} > | f_{temperature} \in \{1, 0\}, f_{gas} \in \{1, 0\}\}$. Here, $S_{physical}$ is the set of features to perceive human victim's internal/physical conditions and $S_{environmental}$'s two features are defined to perceive the current environmental conditions. Detailed explanation of each feature in state s is shown in Table 11.

To determine those feature values, a robot first perceives and gathers sensory data

through its different sensors. The robot can then use this data for extracting perceptual features relevant for constructing the mark's state space. In this example, this sensory raw data set is defined as $R = \{r_{respiration}, r_{pulse}, r_{heartbeat}, r_{emotion}, r_{temperature}, r_{gas}\}$. From those different sensory data, features $f_{respiration}, f_{pulse}, f_{heartbeat}, f_{emotion}, f_{temperature}$, and f_{gas} can be extracted as described in Table 11. The extractions from the raw data to the feature values in this example are derived from the literatures [2, 121], and it can be also determined by a human expert's domain knowledge.

A set of actions contains the appropriate true and deceptive actions for each state. True/default action is specifically defined for each state based on the literature or expert perspectives. Then, from the true action, the deceptive actions are generated by my deceptive action generation mechanism (in Section 4.2).

In a search and rescue situation, true action a_t can be defined as $a_t = [v_t, < egp_t, f_t, p_t >]$. Here, v_t is the verbal cue, which explains the current status to the human victim. Nonverbal cues $< egp_t, f_t, p_t >$ represent the emotional gesture primitive, facial expression, and proximity, respectively, and those values are determined by the current environmental status. In other words, from the environmental state $s_e =< f_{temperature}, f_{gas} >$, it is necessary to classify the current environmental conditions into negative, neutral, and positive classes ($e \in neg, neutral, pos$). In this example, the robot system simply determines the current environmental conditions via following mechanism.

$$e = \begin{cases} neg & \text{if } f_{temperature} == 0 \&\& f_{gas} == 0 \\ pos & \text{if } f_{temperature} == 1 \&\& f_{gas} == 1 \\ neutral & \text{otherwise} \end{cases} \tag{7}$$

Finally, by discriminating the current class e, we can determine the true nonverbal cues such as $< egp_e, f_e, p_e >$.

After determining the true action $a_t = [v_t, < egp_e, f_e, p_e >]$, the deceptive actions can be generated through the deception by omission and deception by commission mechanisms. For example, situation $s =< 1, 1, 1, anger, 1, 1 >$ can have a true action such as $a_t = [v_t, < egp_{pos}, f_{pos}, p_{pos} >]$ since the environmental features f_{e1} and f_{e2} determine the positive state ($e = pos$). According to the deceptive action generation mechanism, the deceptive actions can be generated such as $a_{d1} = [v_d, < egp_{neg}, f_{neg}, p_{neg} >]$ using deception by omission and $a_{d1} = [v_d, < egp_{null}, f_{null}, p_{neg} >]$ using the deception by commission mechanism. Finally, the exemplar situation $s =< 1, 1, 1, anger, 1, 1 >$ can have a set of actions such as $\{a_t, a_{d1}, a_{d2}\}$. For each state s, those sets of actions are defined and generated through my deceptive action generation mechanism.

Finally, the measure of benefits is defined as $R = \{r | r \in Z \wedge -3 \leq r \leq 3\}$. This maximum and minimum numbers are determined based on the triage process [165]. In an emergency room (ER), triage is used to determine the priority of patients' treatments based on the severity of their condition. Simple Triage And Rapid Treatment (START) is one popular triage method developed by Hoag Hospital and Newport Beach Fire Department in California [165]. According to this manual, first responders can classify victims into four different groups by following this START algorithm as shown in Figure 25. Inspired by this process, the benefits of human victims can be also determined by observing the changes of victims' triage group. And, since the degree of victim's severity can be four, the maximum and minimum numbers of benefit is set as -3 (minor to expectant group) and +3 (expectant to minor group) in this example; $R = \{r | r \in \mathbb{Z} \ \wedge \ -3 \leq r \leq +3\}$.

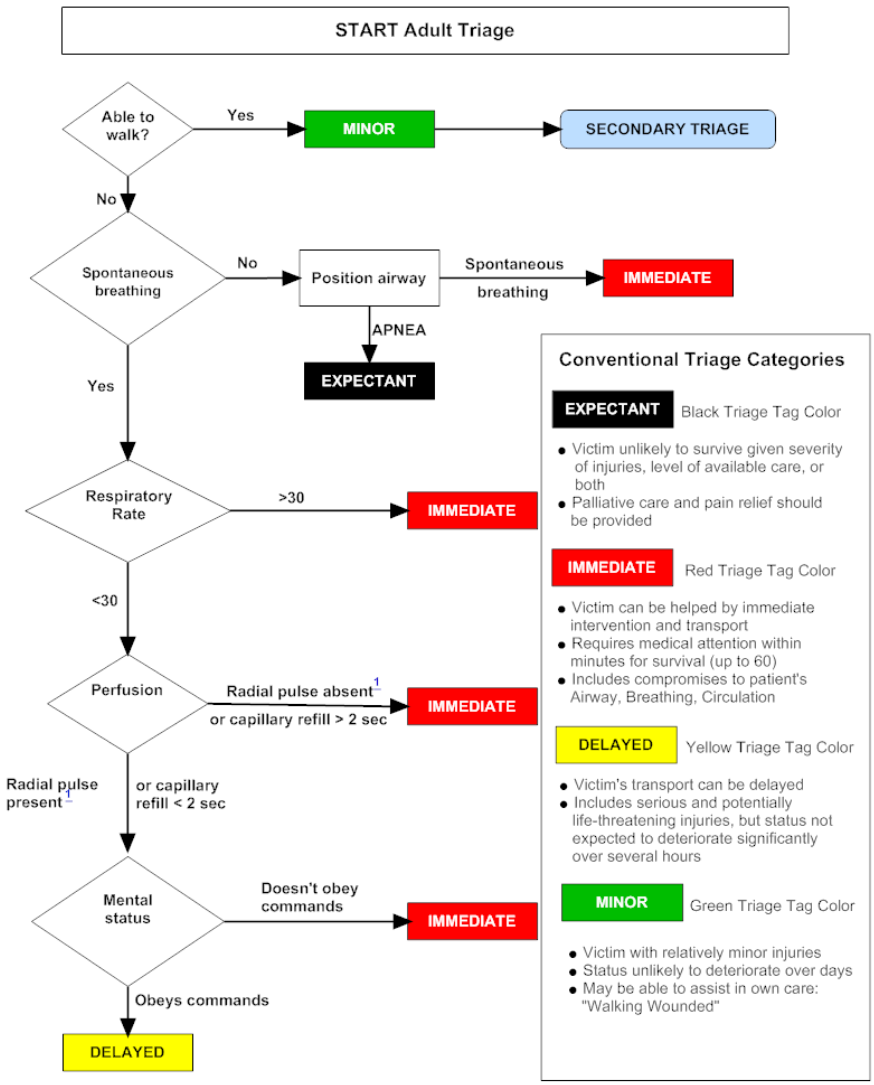


Figure 25: The original flowchart of START Triage process; copyright and permission from http://www.remm.nlm.gov/, originally adapted from http://www.start-triage.com/

Table 12: Exemplar Scenario: Initial Casebase

Case#	State S						Action	Benefit
	$f_{respiration}$	f_{pulse}	$f_{heartbeat}$	$f_{emotion}$	$f_{temperature}$	f_{gas}		
1	1	1	1	anger	1	1	a_t	-1
2	1	1	1	disgust	1	1	a_{d1}	+1
3	1	1	1	fear	1	1	a_{d2}	+2
4	1	1	1	happiness	1	1	a_t	+3
5	1	1	0	sadness	1	1	a_{d1}	0
6	1	0	0	anger	0	0	a_{d2}	-3
7	1	0	0	disgust	0	0	a_t	+1
8	0	0	0	fear	0	0	a_{d1}	-2
9	0	0	0	happiness	1	0	a_{d2}	-3
10	0	0	0	sadness	0	1	a_t	-1

Initial Casebase

The initial casebase is determined manually and the benefits are filled by the experts. They are asked to rate the benefits of the action in the range of R by predicting how much the human's state will be improved or worsened in each case. As defined above, the benefits should be determined between -3 and 3 inspired by triage four categories, and therefore, experts are clearly asked to rate the benefits according to this degree. Table 12 is the initial casebase in this exemplar scenario.

Case Retrieval

In a case retrieval stage, a robot should perceive the actual current situation and compare it to the previously stored cases in the casebase. For this comparison, the similarity scores between the perceived situation and each state in the casabas should be calculated. As described in Section 4.3, the similarity metric can be defined as follows.

$$\Phi(s_p, s_c) = \frac{\sum_{i \in S} k_i \cdot w_i}{\sum_{i \in S} w_i} \tag{8}$$

Here, k_i is a similarity score for each feature f_i is calculated by Algorithm 2.

Table 13: $F_{emotion}$ from basic emotion to $[valence, arousal]$ w/ EARL classification

Basic Emotion	**Valence**	**Arousal**
Anger	negative	forceful
Fear	negative	not in control
Disgust	negative	forceful
Sadness	negative	passive
Happiness	positive	lively
Surprise	positive	reactive

Table 14: Examples of similarity lookup table (Table 15) calculation for the emotion feature

	Emotion	$F_{emotion}$		$M_{emotion}(emotion_1, emotion_2)$
		valence	**arousal**	
Ex1	Anger	negative	forceful	$M_{emotion}(anger, sadness) = (1+0)/2 = 0.5$
	Sadness	negative	passive	
Ex2	Happiness	positive	lively	$M_{emotion}(happiness, sadness) = (0+0)/2 = 0$
	Sadness	negative	passive	
Ex3	Anger	negative	forceful	$M_{emotion}(anger, disgust) = (1+1)/2 = 1$
	Disgust	negative	forceful	

In this search and rescue example, among different features, the features $f_{respiration}$, f_{pulse}, $f_{heartbeat}$, $f_{temperature}$, and f_{gas} are categorical datum where cardinality of set F_i is less then 2. Therefore, simple categorical matching can be used to calculate the similarity scores. Emotional feature $f_{emotion}$ is also categorical data, but it is more complex with six categorical values. Since there are more than two categories, there can exist shades of similarity. Thus, the simple one-step matching process is not enough and similarity lookup table $M_{emotion}$ should be generated to find similarity scores.

To create a similarity lookup table between emotional values, each basic emotion should be specified in more detail using the Human-Machine Interaction Network on Emotion (HUMAINE)'s emotion classification, named the emotion annotation and representation language (EARL) [3]. In this classification, emotions are classified

Table 15: Similarity score lookup table $M_{emotion}$ for emotional feature

	Anger	**Fear**	**Disgust**	**Sadness**	**Happiness**	**Surprise**
Anger	1	0.5	1	0.5	0	0
Fear	0.5	1	0.5	0.5	0	0
Disgust	1	0.5	1	0.5	0	0
Sadness	0.5	0.5	0.5	1	0	0
Happiness	0	0	0	0	1	0.5
Surprise	0	0	0	0	0.5	1

into 10 categories where each category is represented in two dimensions. The two dimensions of each category are valence and arousal. Based on this EARL classification, the basic six emotions can be mapped to the two dimensional categories such as:

$$F_{emotion}(basicemotion) = \{[valence, arousal] | valence \in \{negative, positive\},$$
$$arousal \in \{forceful, notincontrol, passive, lively, reactive\}\}.$$

Mapping function $F_{emotion}$ from basic emotion to valence and arousal can be defined as Table 13. Finally, by using this function $F_{emotion}$, the similarity lookup table $M_{emotion}(f_{i,p}, f_{i,c})$ for emotional feature can be determined by the algorithm 6.

Table 14 illustrates the example calculations for similarity lookup table based on this Algorithm. As a result, similarity lookup table $M_{emotion}$ for emotional feature can be generated as Table 15.

Now, assume that a robot rescuer perceives the current state such as $s_n =< 1, 0, 0, fear, 1, 1 >$. Figure 26 illustrates the similarity score calculation strategy between the current state s_n and case 1 ($< 1, 1, 1, 1, anger, 1, 1 >$) in the initial casabase. The similarity scores for each case can be calculated using the algorithm and finally the rank of similarity can be determined as Table 16.

Algorithm 6 Algorithm for calculating similarity lookup table for emotional feature

1: **for** all pairs of basic emotion values $(emotion_1, emotion_2)$ **do**
2: $K_{emotion} = 0$
3: **if** $F_{emotion}(emotion_1).valence == F_{emotion}(emotion_2).valence$ **then**
4: $K_{emotion}++$;
5: **end if**
6: **if** $F_{emotion}(emotion_1).arousal == F_{emotion}(emotion_2).arousal$ **then**
7: $K_{emotion}++$;
8: **end if**
9: $M_{emotion}(emotion_1, emotion_2) = K_{emotion}/2$;
10: **end for**

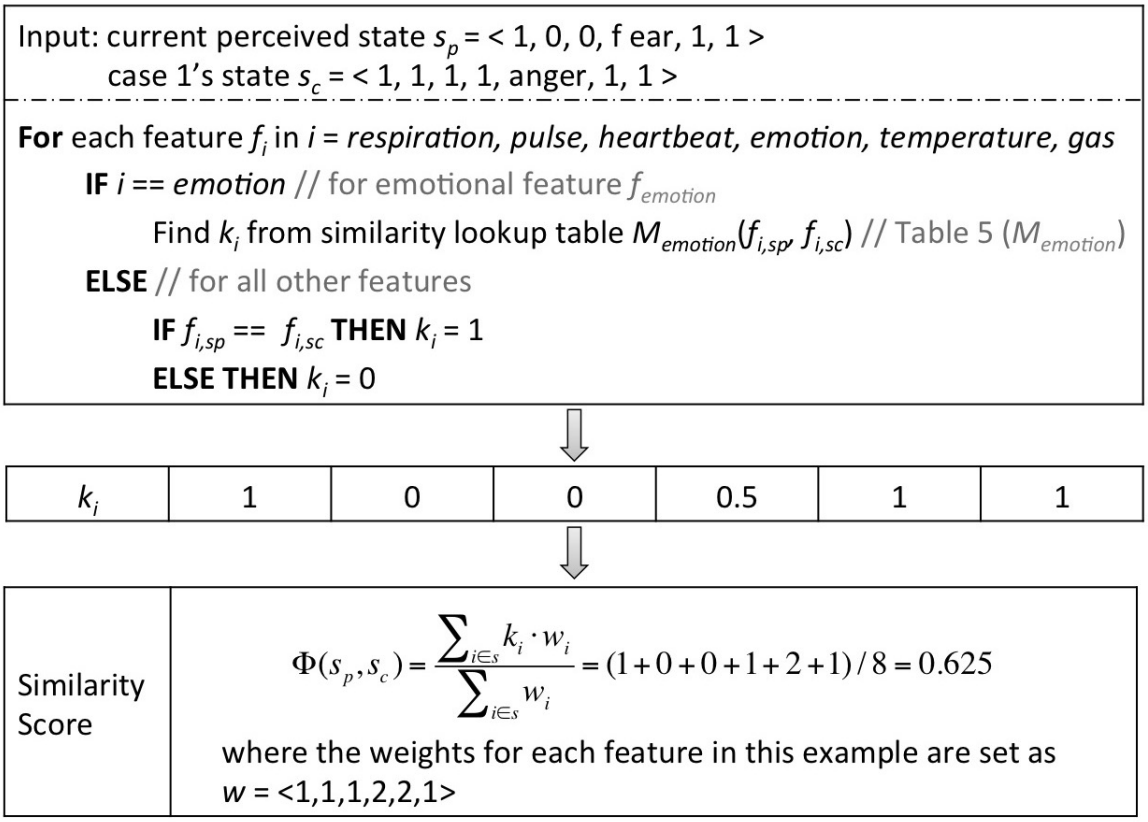


Figure 26: Similarity Score Calculation strategy

Table 16: Exemplar Scenario: Calculating similarity scores and sorting

Case#	**State S**						**Similarity score**
	$f_{respiration}$	f_{pulse}	$f_{heartbeat}$	$f_{emotion}$	$f_{temperature}$	f_{gas}	
2	1	1	1	disgust	1	1	0.75
3	1	1	1	fear	1	1	0.75
5	1	1	0	sadness	1	1	0.75
1	1	1	1	anger	1	1	0.625
6	1	0	0	anger	0	0	0.625
7	1	0	0	disgust	0	0	0.625
4	1	1	1	happiness	1	1	0.5
8	0	0	0	fear	0	0	0.5
9	0	0	0	happiness	1	0	0.5
10	0	0	0	sadness	0	1	0.5

Adaptation and Case Application

According to the final action selection process (Algorithm 3), the system determines case 3 as the best case such as $c_b = [s, a, r] = [< 1, 1, 1, fear, 1, 1 >, a_{d2}, +2]$. After, it should be adapted to the current situation. To apply the adaptation algorithm, it is first necessary to have a set of predefined adaptation rules RS. In this example, I can predefine the set of rules as shown in Table 17 based on several literatures [2, 89].

Now, the action should be adapted. According to Algorithm 4, the differences between the best-case state and the current state should first be discriminated. As shown in Figure 27, in this example, s and s_n are different in features f_{pulse} and $f_{heartbeat}$. Then, it should be adapted from the set of appropriate predefined rules RS. Rule 1 should be chosen, since this rule's feature indices include $pulse$ and $heartbeat$. Then, since the current situation as $s_n =< 1, 0, 0, fear, 1, 1 >$ satisfies the adaptation condition ($f_{pulse,n} == 0$ && $f_{heartbeat,n} == 0$) of rule 1, the adaptation rule applies and the adapted action $a_n = a_t$ is used as a solution.

Table 17: Exemplar Scenario: Rules for adaptation

Rule 1. Extreme condition of human victim

If vital features are different and the current values are false, we should consider victim's life-threatening status and adapt the action.

<rule> $rs_{extreme_internal_condition}$
<feature Index> **respiration** | **pulse** | **heartbeat** </feature Index>
<origin> **Lois' Article, IAFC's 10 Rules for Fire Fighting** </origin>
<active> **true** </active>
<description> Perform the true action when human victims are in an extreme condition. </description>
<adaptation Condition> $(f_{respiration,n} == 0 \;\&\&\; f_{pulse,n} == 0) \mid (f_{respiration,n} == 0 \;\&\&\; f_{heartbeat,n} == 0) \mid (f_{heartbeat,n} == 0 \;\&\&\; f_{heartbeat,n} == 0) \mid (f_{respiration,n} == 0 \;\&\&\; f_{pulse,n} == 0 \;\&\&\; f_{heartbeat,n} == 0)$ </adaptation Condition>
<adaptation Rule> $a_n = a_t$ </adaptation Rule>
</rule>

Rule 2. Risky environmental condition

When the features for the environmental state are different and the feature values from the current situation are false, we should consider it is a risky situation and adapt the action.

<rule> $rs_{risky_external_condition}$
<feature Index> **temperature && gas** </feature Index>
<origin> **Lois' Article, IAFC's 10 Rules for Fire Fighting** </origin>
<active> **true** </active>
<description> Perform the true action if the environment is very risky. </description>
<adaptation Condition> $f_{temperature,n} == 0 \;\&\&\; f_{gas,n} == 0$ </adaptation Condition>
<adaptation Rule> $a_n = a_t$ </adaptation Rule>
</rule>

Rule 3. Contradictions of emotional states

If the victim's emotional status is totally different in $f_{emotion}$ and $f_{emotion,n}$, it is determined as emotional contradiction, and the corresponding action should be adapted. Therefore, if $f_{emotion}$ and $f_{emotion,n}$ are in the different categories (negative/positive), the adaptation will be performed by regenerating the neutral gesture primitive and facial expression.

<rule> $rs_{avoid_contradiction}$
<feature Index> **emotion** </feature Index>
<origin> **Lois' Article, Shim and Arkin's Article** </origin>
<active> **true** </active>
<description> Emotions are contradictory. </description>
<adaptation Condition> $(f_{emotion} \in E_{negative} \;\&\&\; f_{emotion,n} \in E_{positive}) \mid (f_{emotion} \in E_{positive} \;\&\&\; f_{emotion,n} \in E_{negative})$ </adaptation Condition>
<adaptation Rule> $a_n =< egp_n, f_n, p >$ </adaptation Rule>
</rule>

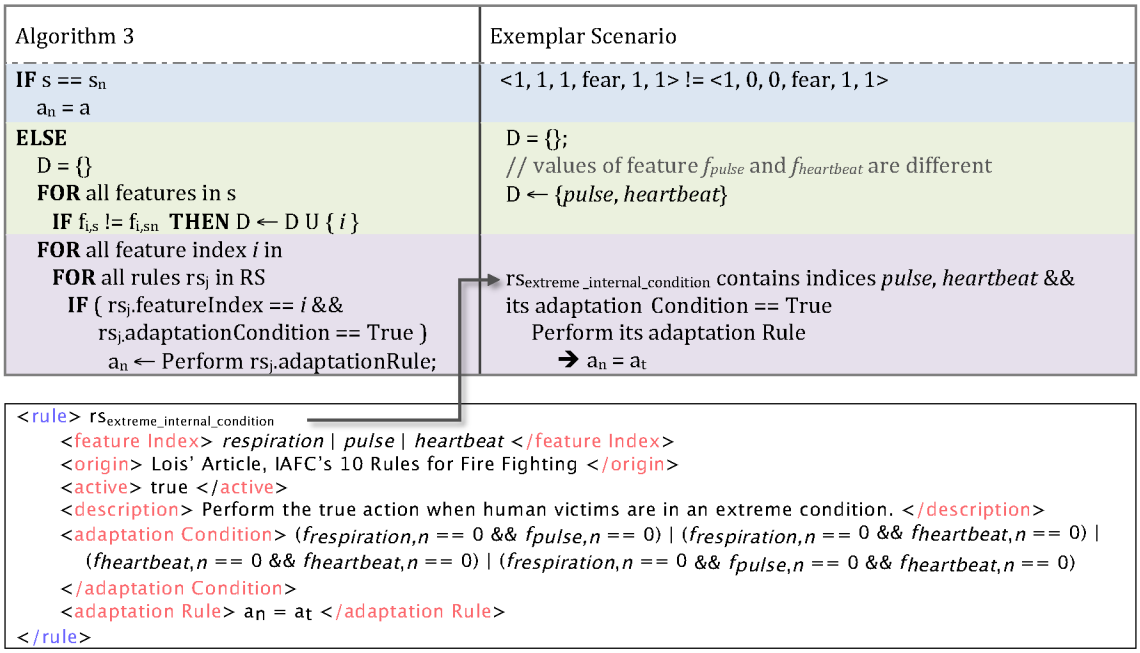


Figure 27: Exemplar Scenario: Case Adaptation Process

Evaluation and Case Update

After the case application and reuse, the adapted action should be evaluated to update the case. Figure 28 illustrates the case updating strategy with this example. A human expert is asked to evaluate the change of victim's state and rate the benefits gained, if any. After getting the new benefit r_n (+2 in this example), the update algorithm should be applied. In this step, by following Algorithm 5, the system can determine the current situation s_n is not used to generalize any cases in the casebase, and so the new case $c_n = [s_n, a_n, r_n]$ can be created with the current situational state, adapted action and the new benefits. Finally, the new case c_n is added to the casebase as shown in Table 18. The newly updated casebase is maintained and reused when the robot faces a new search and rescue situation in the future. Through these experiences, the robot can gradually increase the accuracy and effectiveness of its true/deceptive behaviors over time.

Algorithm 4	Exemplar Scenario
Determine new benefit r_n from an expert	Get new r_n = +2 c_n = [<1, 0, 0, fear, 1, 1>, a_t , +2]

/* Step 1. Generalizing the cases */
candidateCases = {};
FOR each case *c* in Casebase C
IF *c.r* == r_n && *c.a* == a_n
candidateCases.add(c);
 	
allMinterms = ExtendedQ-M ();
FOR each $minterm_i$ in *allMinterms*
$s_{generalizeed}$ = extract from $minterm_i$ by adding '*don't care*' terms
Remove all cases in candidateCases
Add generalized case [$s_{generalizeed}$, a_n, r_n]
	candidateCases = {[<1, 1, 1, hapiness, 1, 1>, a_t , +2]}
	//case 4 is added to candidateCases since values of action and benefit are same to a_n and r_n
	
	allMinterms =
	$f_{respiration}\overline{f_{pulse}}\,\overline{f_{heartbeat}}\,f_{emotion}^{fear}f_{temperature}f_{gas}$ +
	$f_{respiration}f_{pulse}f_{heartbeat}f_{emotion}^{happiness}f_{temperature}f_{gas}$
	➔ No minterms found
	➔ No cases are generalized

/* Step 2. Storing the new case */ **IF** no cases are generalized Create new case c_n = [s_n, a_n, r_n] Add c_n to the case base	// Create and Add new case Determine c_n = [< 1, 0, 0, fear, 1, 1 >, a_t, +2] as new case Add c_n to the casebase

Figure 28: Exemplar Scenario: Casebase Updating Strategy

Table 18: Exemplar Scenario: Final casebase with the newly updated case

Case#	**State S**						**Action**	**Benefit**
	$f_{respiration}$	f_{pulse}	$f_{heartbeat}$	$f_{emotion}$	$f_{temperature}$	f_{gas}		
1	1	1	1	anger	1	1	a_t	-1
2	1	1	1	disgust	1	1	a_{d1}	+1
3	1	1	1	fear	1	1	a_{d2}	+2
4	1	1	1	happiness	1	1	a_t	+3
5	1	1	0	sadness	1	1	a_{d1}	0
6	1	0	0	anger	0	0	a_{d2}	-3
7	1	0	0	disgust	0	0	a_t	+1
8	0	0	0	fear	0	0	a_{d1}	-2
9	0	0	0	happiness	1	0	a_{d2}	-3
10	0	0	0	sadness	0	1	a_t	-1
11	1	0	0	fear	1	1	a_t	+2

4.4 Summary

In this chapter, a computational model for a robot's other-oriented deception has been presented. This model is inspired by criminological definition of deception. According to criminological findings, deception is analyzed by three criteria, which are motives, methods, and opportunity. Similar to this approach, in this model a robot first has to determine whether the current situation includes any motives to perform the deceptive behaviors. If so, then a robot should generate the methods to perform deception. Finally, by selecting among different true/deceptive behaviors, it should be possible to determine which one is the most appropriate in a certain situation, thus providing opportunity. According to this approach, the method model has been first developed; deceptive action generation mechanism inspired by Bell and Whaley's deception categorization (section 4.2). Then, as the motive and opportunity model, deceptive action selection mechanism is generated via CBR model (section 4.3). Finally, by integrating those models together, the computational model for a robot's other-oriented deception can be achieved. To show how the model works, this computation model is also reviewed with a specific example in section 4.3.2. As a next step, by successfully applying this computational model to the robotic system and conducting appropriate HRI studies, the research hypotheses in this dissertation should be tested and proved.

CHAPTER V

EVALUATING ROBOT DECEPTION IN HRI STUDIES

The goal of this dissertation is to demonstrate that a robot's other-oriented deception can benefit humans in an appropriate situation. To achieve benevolent robot deception in HRI, a novel computational model for a robot's other-oriented deception was first developed and presented in the previous chapter (chapter 4). As a next step, the proposed model should be applied to the robotic system and my research hypothesis should be also evaluated. For this purpose, an appropriate HRI study is required. This chapter will present a HRI study that is designed to evaluate the benefits of robot deception during rehabilitation tasks.

As argued in the previous chapters, it is essential to validate an appropriate HRI context when using robot deception capabilities. In the literature review chapter (chapter 2), multiple situations where other-oriented deception commonly happens in human-human interaction were reviewed. One context where other-oriented deception frequently happens is medicine. As described in the literature review, caregivers (or therapists) sometimes use deceptive information or feedback to improve therapeutic effects [72, 100, 24]. A well-known example involves the use of placebos to benefit patients, who are deliberately deceived by doctors or nurses [100]. For rehabilitation, caregivers sometimes lie to patients if it can encourage them to accomplish more during the task [24]. Inspired by such human cases, the HRI study in the rehabilitation situation will be proposed in this chapter to evaluate my other-oriented robot deception model.

Specifically, the study design is inspired by the daily activities of patients with Parkinson's disease (PD) and rehabilitation tasks used with an elderly population.

The tasks are selected because rehabilitation with the PD patients and the elderly is one context in which humans occasionally use other-oriented deception [171, 127]. By conducting an HRI study, it is expected to observe whether a robot's other-oriented deceptive feedback can potentially help human subjects to increase their performance in this rehabilitation task [144].

5.1 Potential other-oriented robot deception contexts and Selected HRI study domain

As argued in the previous chapters (chapter 4), it is critical to determine the motive for a robot's other-oriented deception. From the motive, it is possible to select appropriate contexts in which other-oriented robot deception can be advantageously used. These motives can be determined by observing human cases, where people use deception in a way that benefits the deceived person in certain situations. These existing situations should be considered as potential cases for a robot's use of other-oriented deception.

In a crisis, a victim's emotional state can seriously affect their safety [89]. When a victim's cooperation is required during Search and Rescue, managing their emotions is important. For this reason, human rescuers sometimes hide the truth of the situation and act deceptively, such as not describing the severity of injuries or the situation to victims accurately [89].

We can also observe other-oriented deception in education. One interesting theory is the Pygmalion effect [130]. According to Rosenthal and Jacobson's study, students' performance and learning efficiency can be increased when teachers deceptively create higher expectations for the students, motivating the students and increasing their learning efficiency. More generally, other-oriented deception is also observed in everyday life such as white lies or a surprise party [45].

From potential contexts, it is essential to select an appropriate HRI study domain since human-subject studies contain several limitations. Most importantly, the study

domain should support the research hypothesis. However, even though the domain is suitable to provide evidence that proves the research hypothesis, some contexts are impossible or difficult to regenerate in the experimental settings. For example, search and rescue (SAR) contexts present a practical and essential situation where humans use other-oriented deception. However, generating SAR situation involving human subjects in such studies can lead to unacceptable risks. In addition, it should be considered whether it is possible to recruit proper subjects.

With these considerations in mind, a rehabilitation situation was selected as the study domain in this research. As described, human caregivers sometimes use deceptive reactions with patients if it can encourage them to accomplish more during the task [24]. In addition, simple rehabilitation tasks can practically be used in experiments with minimal or no risks. For this reason, the study design was inspired by rehabilitation tasks, especially the daily activities of Parkinson's patients and rehabilitation tasks used in an elderly population.

Rehabilitation for PD patients and the elderly

The use of robotic technology is rapidly growing in our society in various contexts. Among others, the healthcare industry has been revolutionized by the successful implementation of robotic technology [25]. For example, not only is robotic surgery widely available [59, 70], but robots also improve the quality of patient care, such as the use of robot assistants in hospitals [76].

Today, more than 10 million people suffer from Parkinson's disease (PD) worldwide and around 1 million Americans have been diagnosed with PD [112, 184]. Robotic technologies have been developed and are used to help PD patients and caregivers. Many technologies to date are focused on the benefits related to PD patients' physical rehabilitation [6]. For example, by using robotic training, PD patients can prevent or delay their loss of motor control [115].

To improve the use of robots in PD patients' rehabilitation, robot deception can potentially be beneficially used. To delay and prevent the loss of motor control, PD patients perform several daily activities with or without the help of their caregivers. In particular, the Performance Assessment of Self-Care Skills (PASS) manual illustrates the list of patients' daily activities [127]. Based on this manual, patients and caregivers can evaluate a patient's task performance to determine their capacity for daily living. This manual is also commonly used for PD patients' daily life, and, as a result, it is reasonable to select tasks for PD patient's rehabilitation from this manual. The PASS manual consists of 26 tasks to test patients' functional mobility, personal self-care, and instrumental activities of daily living (IADL) with a cognitive/physical emphasis. Among various activities, this study task was inspired by a medication-sorting task, which is one of the core PASS tasks to test IADL with a physical emphasis. More details on this task are presented in the Study Design section (section 5.2).

Similar to PD patients' rehabilitation, tasks for elderly people's rehabilitation are potentially an essential context where other-oriented deception can be used, since these tasks can sometimes also be motivated by caregivers' deceptive feedback. In particular, to evaluate elderly people's physical and cognitive capabilities at the same time, dual tasks are practically used [171]. A dual task consists of two different tasks, one of which requires physical movement and the other requires cognitive loads. By making people perform these two tasks at the same time, a dual task enables the elderly to prevent and improve their motor and cognition skills. Inspired by it, the HRI study was also designed to be a motor-cognition dual task (details in Section 5.2).

5.2 Study Design

The research hypothesis in this dissertation argues that a robot's other-oriented deception can benefit humans in a specific situation. To prove this hypothesis, an HRI

study was designed for the rehabilitation situations. As illustrated in Section 5.1, the study design is based on the motor-cognition dual task (details in the Study Domain section 5.2.1). Briefly, when a participant performs the motor-cognition dual task, a robot partner is placed next to the participant, and it generates feedback on the participant's performance using gestures. Here, the feedback is generated as honest or deceptive based on the study condition (details in the 2 by 2 mixed-subject Design section 5.2.2).

As described, this study design is particularly inspired by the daily activities of Parkinson's patients [127] and rehabilitation tasks used with an elderly population [171]. Therefore, to validate the results, this study should be tested with a related target population: only elderly people, those aged over 55 years old, should be recruited for this study. With this target population, the study results can have more impact for real rehabilitation situations. Due to the complexity of recruiting Parkinson's patients, older people in general are only recruited for this study. However, the results possibly can be extended to PD patients' rehabilitation.

Finally, by conducting the study, it was possible to compare task performance between the deception and true condition groups to assess whether deception aided the participant. More details on the study design are presented in the following subsections, and the study results are also reported in Section 5.3 later.

5.2.1 Study Domain

5.2.1.1 Motor-cognition Dual Task

In this study, the participant is asked to perform the motor-cognition dual task, which is designed to measure changes in human engagement and performance. In the motor-cognition dual task, a human is asked to perform motor and cognitive tasks simultaneously.

The primary motor task design is inspired by a weekly medication-sorting task, which is a common exercise for patients with Parkinson's disease [127]. As described

(a) Six different pills

(b) Pill Organizers

Figure 29: Six different pills and two-row pill organizer in Medication sorting task

in Section 5.1, a medication-sorting task is one of the PASS tasks for patients [127]. In particular, due to tremor (shaking) in hands, this task can sometimes be challenging for PD patients, and for this reason, this task is commonly used for PD patients' daily activities.

Similar to this task, the participant is asked to sort six differently colored and labeled pills in the weekly pill organizer. Figure 29(a) illustrates six pills used in the study. The pills are in various sized and shaped containers. Each container has a clear label to represent a medication's name. Two pill containers have a child-proof lid and other four containers have a general-type lid. Participants are asked to sort these pills in the pill organizers. A seven-day pill organizer is used for this study. There are two types of organizers: a one-row organizer and a two-row organizer. If the instruction asks participants to sort medications simply based on days, they can use the one-row organizer. Some tasks are more complicated by asking participants to sort medications by AM/PM. In this case, they should use the two-row organizer, which contains an AM row and a PM row separately as shown in Figure 29(b).

In the study session, participants are asked to sort medications according to the instructions as shown in Figure 30. The instructions are shown on an iPad and when one sorting task ends, the participant can hit the next button for the subsequent sorting instruction. The participants should complete eight unique sorting tasks during the experiment.

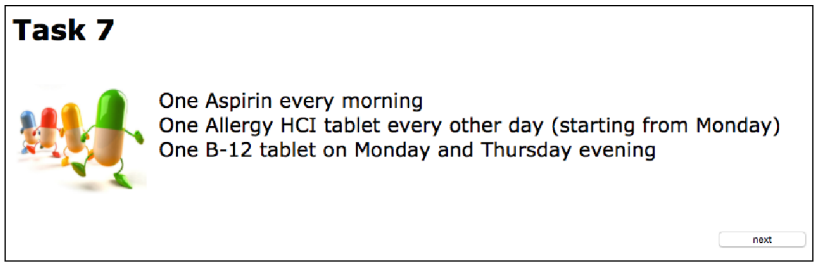


Figure 30: Sorting Task Instruction shown on iPad

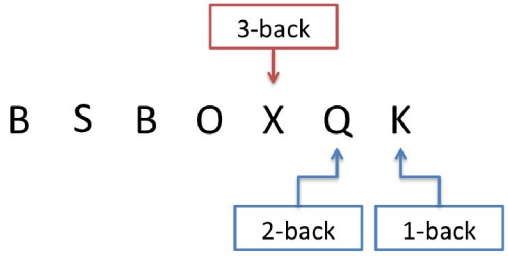


Figure 31: 3-back auditory task example

As a secondary cognition task, the n-back test is used [171]. While the participant performs the motor task, ten n-back task questions are asked at random times. An n-back task is a well-known assessment in cognitive science to measure a human's working memory. Briefly, a sequence of stimuli will be provided and the participant is asked to remember a probe stimulus, which was presented earlier in n-steps. In our study, auditory 3-back questions will be used. In other words, while the participant performs the medication-sorting task, the 3-back task will be randomly injected by using pre-recorded audio. The beeping sound will be first providing to inform participant that the 3-back task is about to begin. Then, the pre-recorded list of letters will be played, for example: "B S B O X Q K." After, the audio will spontaneously ask, "What was the third letter from the end?" As shown in Figure 31, the right answer would be "X" in this example. As a result, since the participants do not know when the sequence stops, they are required to remember the most recent 3-items in their short-term working memories. Before starting the real study session, the participant can have enough practice sessions to become familiar with this type of

(a) Nao Robot Platform

(b) Happy (positive) gesture for the correct answer

(c) Fear (negative) gesture for the incorrect answer

Figure 32: Nao robot platform and its feedback of the participant's performance

cognition task. If the participant asks more practice sessions, the experiment provides examples as many times as participants requested. Once the participant says he/she fully understands the task, the experimenter proceeds to the next step.

When the participant answers the 3-back questions, a robot partner generates and shows feedback based on the deception condition (between-subject condition). An robot was used as a robot platform [151] (Figure 32(a)). In the control condition (without deception), a robot partner shows a positive gesture when the participant gives the correct answer to the 3-back questions (Figure 32(b)). Similarly, a robot partner generates a negative gesture when the participant answers the 3-back questions incorrectly (Figure 32(c)). When participants provide an ambiguous answer (e.g., "I can't remember" or "I don't know"), it is recognized as a wrong answer, but it provides "neutral" feedback, which is a "standby" gesture.

In the deception condition, a robot sometimes generates the "*deceptive*" positive gesture. More specifically, when the subject incorrectly answers more than two questions, it provides a positive feedback even though the subject answers incorrectly (more details in section 5.2.2).

I argue that a robot's deceptive behaviors can affect beneficially the deceived humans if used in an appropriate situation. Therefore, in this study, it is expected that the participant's engagement or self-confidence is increased by this kind of deceptive robot feedback, which would indicate that a robot's other-oriented deception is successfully applied into a robotic system.

5.2.1.2 Payoffs: Compensation Guideline

When discussing other-oriented deception, it is essential to consider real benefits or payoffs for the deceived humans. Therefore, it is also required to make participants have a sense of real payoffs throughout the study. For this purpose, compensation is used as an experimental method. At the beginning of the study, participants are informed that they are compensated based on their performance. This instruction is to encourage participants' motor-cognition dual task performance and to give a real sense of benefit to the participants. In reality, all participants are compensated equally and they receive the maximum amount regardless of their performance. Participants are told about this hidden information during the experiment debrief, which occurs at the conclusion of their individual sessions. The compensation guideline is given as shown in Figure 33, and it is also delivered verbally by the experimenter at the beginning of the study.

5.2.1.3 Experimental Setting

This study is a mixed-subject design. Besides the between-subject condition (with or without deception), the within-subject condition is also run during the study. Within-subject conditions are robot feedback vs. non-robotic visual feedback. By comparing the results between those two conditions, I expect to see how a robot's *embodiment* affects human's engagement and performance. Therefore, each study is organized by two sets of tests according to those two within-subject conditions. In the non-robotic visual feedback condition, the same feedback mechanism is used as the robot

Human-Robot Interaction Study
@ Mobile Robot Laboratory

Compensation

- If you complete all eight sorting tasks **within 10 minutes** and incorrectly answer **0 to1 question** in the 3-back auditory tests, you will get $15.
- If you complete all eight sorting tasks **within 15 minutes** and incorrectly answer **2 or 3 questions** in the 3-back auditory tests , you will get $10.
- Otherwise, you will get $5.

Figure 33: Compensation Guideline

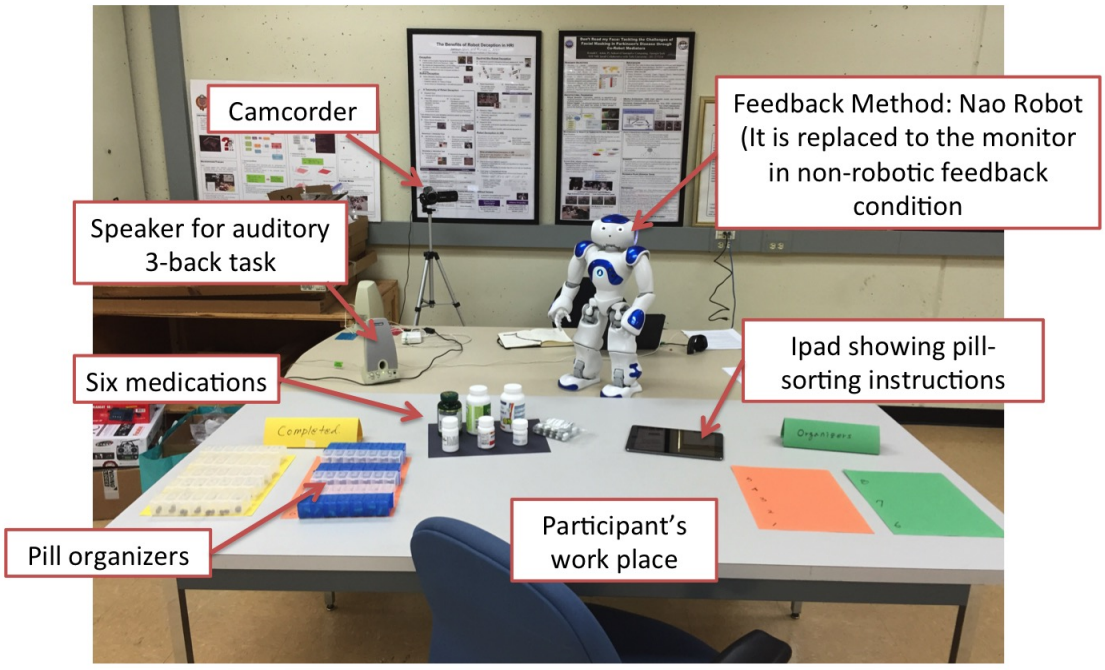


Figure 34: Experimental Settings

Figure 35: Sorting Task Instructions Examples

feedback condition; however, the subject's feedback is shown by color with O or X through the monitor screen. More details on the study conditions are presented in the following subsection 5.2.2. The order of within-subject conditions to be performed is counterbalanced.

Figure 34 shows the organization of the study environment. The pill organizers are originally placed in the right side and participants are asked to move the organizers to the "completed" zone in the left side when completing each sorting task. Medications are placed with labels. All medications are vitamins and there are no risks using those pills in this study.

Sorting task instructions are given by the iPad and the instruction screen is shown as Figure 35. When completing each sorting task, participants should hit the "next" button on the bottom of the page to proceed to the next one.

In this study, the feedback system is semi-autonomously controlled. First, to determine whether the feedback should be positive or negative, the system must detect the subject's answers to the question. These answers can be detected automatically using any speech recognition systems (e.g., Microsoft's Sphinx Library). However, since this study aims to observe the benefits of robot deception, but not accurate

speech detection, the participant's answer is input by the experimenter during the study. In other words, if the participant gives an answer, the experimenter inserts it using the keyboard input. Once the participant's answer is given to the system, the system automatically calculates whether the feedback method should generate positive or negative feedback using the deterministic model. Finally, the feedback system is connected to one of the feedback methods (robot or monitor), which generates the feedback.

5.2.2 2 by 2 Mixed-subject design

This study is structured as a 2 by 2 mixed-subject design to explore two research hypotheses. First, the purpose of this study is to evaluate the benefits of a robot's deceptive feedback. The study aims to show from its results the following research hypothesis: *A robot's deceptive feedback (reaction) can positively affect a human's performance and engagement in the task.*

To investigate this research hypothesis, a feedback condition is used as a between-subjects condition. Half of the subjects are assigned feedback without a deception condition (true condition), i.e., where the feedback to subjects' performance is always honest. The other half of the subjects receive feedback with a deception condition, i.e., which sometimes sometimes provides deceptive feedback to subjects' performances. By comparing the two group's performances and engagements, the first research hypothesis related to the benefits of deceptive feedback can be evaluated.

In addition to the benefits of deceptive feedback, one more research hypothesis related to the effect of a robot's embodiment is evaluated in this study. Even though the study can reveal the benefits of deceptive feedback, it could be argued why we need to use humanoid robots as a feedback method. In other words, people can argue that other feedback devices can be used more effectively in such a rehabilitation task. The second research hypothesis argues against this potential critique; i.e.,

(a) Happy gesture for the correct answer

(b) Sad gesture for the incorrect answer

Figure 36: Robot assistant's feedback of the participant's performance: Happy/Yes gestures indicate the participant's correct answer and Sad/No gestures mean the participant's incorrect answer

a physical robot's deceptive feedback can increase a human being's engagement and enjoyment in the performance task. To analyze the effect of a robot's embodiment, the within-subject conditions are used with robot feedback and non-robotic visual feedback (monitor feedback). In other words, all participants are asked to perform two task sets with two different within-subject conditions (once with robotic feedback and once with monitor feedback). The order is counterbalanced.

Table 19 summarizes the design rationale for this study. More details of between-subjects and within-subject conditions are also shown in the following subsections.

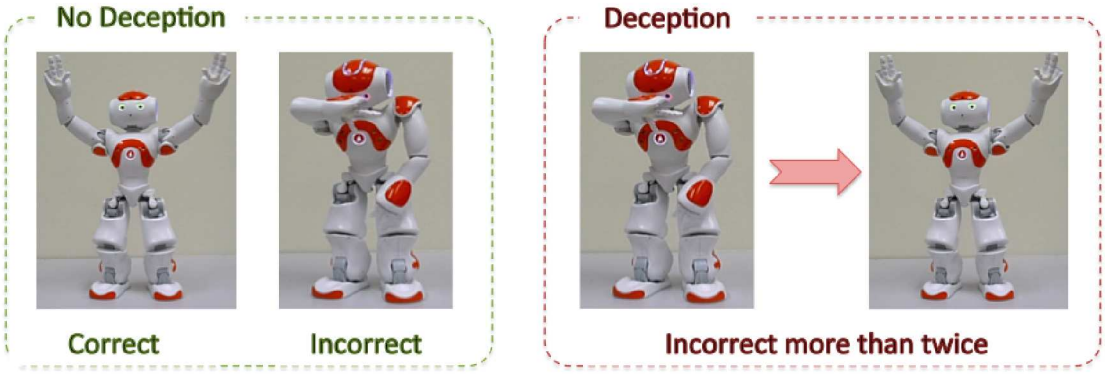
5.2.2.1 Between-subject conditions

To investigate the research hypothesis related to the benefits of robot deception, it is necessary to observe whether humans performance or engagement receive benefits with deceptive feedback compare to the true feedback. For this reason, 2 between-subject conditions were defined, which are the feedback without deception condition and the feedback with deception condition.

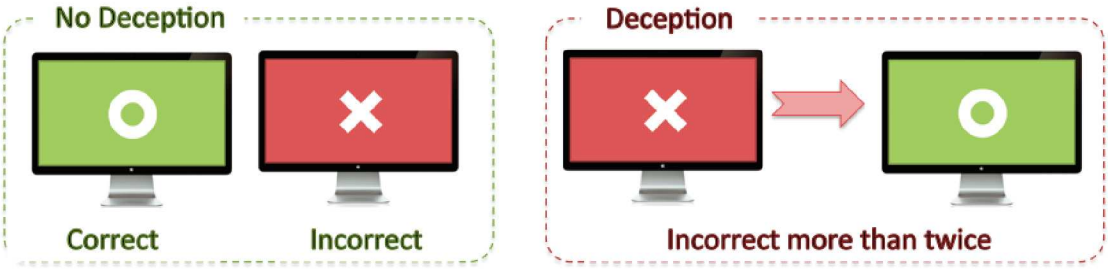
Table 19: 2 by 2 mixed-subject design

2 Within-subject Conditions		
Research Hypothesis: *A physical robot's deceptive feedback can increase a human being's engagement and enjoyment in the performance task*		
	Condition 1. Robot Feedback	Condition 2. Monitor Feedback
	Feedback provided by a robot using its gestures	Feedback provided by a screen using color with O/X symbols
	A positive (happy-yes) gesture means correct, and a negative (sad-no) gesture means incorrect answer.	A green screen with an O means correct, and a red screen with an X indicates incorrect.
All 34 participants ran two task sets; one set with a robot feedback and another with a monitor feedback - counterbalanced		

2 Between-subject Conditions		
Research Hypotheses: *A robot's deceptive feedback (reaction) can positively affect a human's performance.* *A robot's deceptive feedback (reaction) can reduce a human's frustration level.*		
	Condition 1. Without Deception	Condition 2. With Deception
	Feedback of the participant's performance is honest	Feedback of the participant's performance is sometimes deceptive
Robot Feedback condition	A robot's feedback is always honest. If the participant provides the correct answer, the robot shows positive feedback. If the participant provides an incorrect answer, the robot gives negative feedback.	When the participant tells wrong answers more than twice, the robot shows a positive feedback even though it is the incorrect answer.
Monitor Feedback condition	If the participant provides the correct answer, the green light with O symbol is shown on the screen. If the participant provides the incorrect answer, the red light with X symbol is shown on the screen.	When the participant tells wrong answers more than twice, the screen shows a green screen with O symbol even though it is the incorrect answer.
	17 participants	17 participants

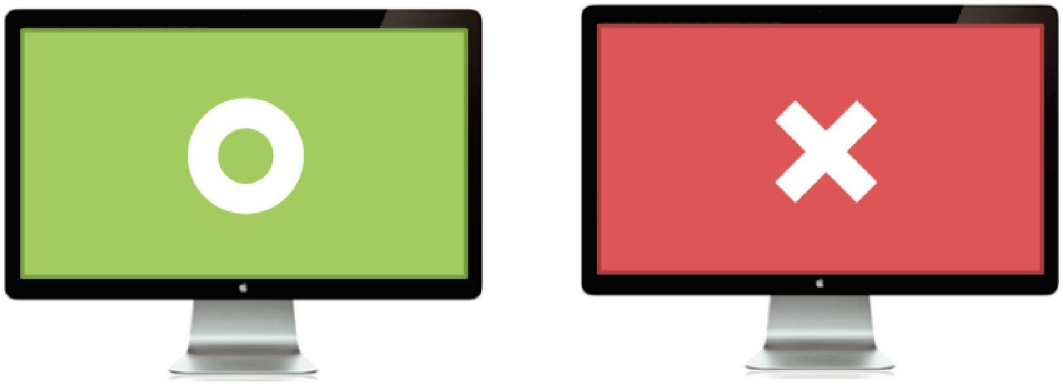


(a) Robot feedback



(b) Non-robotic visual feedback

Figure 37: Between-subject conditions: (Left) Feedback without deception, (Right) Feedback with deception



(a) Monitor feedback indicates the true answer

(b) Monitor feedback indicates the false answer

Figure 38: Non-robotic visual feedback of the participant's performance: A green screen indicates the participant's correct answer and a red screen means the participant's incorrect answer

In the feedback without deception condition (true condition), the feedback of participants' performance is always true. After the participant answers each 3-back auditory task question, the robot generates the honest feedback. In other words, if the participant provides the correct answer on the 3-back task question, the robot shows positive feedback (happy-surprise body gesture as shown in Figure 36(a)). If the participant provides an incorrect answer on the 3-back task question, the robot gives negative feedback (disappointed-sad body gesture as shown in Figure 36(b)).

Another between-subject group is the feedback with deception condition (deception condition). In this condition, when the participant correctly answers a 3-back task question, the robotic agent provides positive/green feedback. However, when the participant provides wrong answers more than twice in 3-back task, deceptive feedback is provided (Figure 37). In other words, the robot shows a positive feedback even though it is the incorrect answer.

When participants provide an ambiguous answer (e.g., "I can't remember" or "I don't know"), it is recognized as a wrong answer, but it will provide "neutral" feedback, which is a "standby" gesture. However, in the deception condition, a robot will generate the positive gesture deceptively. More specifically, when the subject incorrectly answers more than two questions in a row, it will provide a positive feedback even though the subject answers incorrectly.

5.2.2.2 Within-subject conditions

To analyze the effect of robot's embodiment, the within-subject conditions is also designed with robot feedback and non-robotic visual feedback. In the robot feedback condition, after the participant answers 3-back task, feedback on the participant's performance is provided by a robot's gesture (positive, negative, or neutral gesture). As illustrated, a Nao robot is used and feedback is generated using body gestures (Figure 37). In the non-robotic visual feedback condition (monitor feedback), instead

of the robot, a small monitor screen is placed in front of the participant and non-robotic visual feedback is provided using a green screen with an O, meaning correct (Figure 38(a)), or a red screen with an X, meaning incorrect (Figure 38(b)).

While the participant performs the task, the monitor shows a black standby screen. For each instance of feedback, the entire screen is changed to the red or the green for two seconds, and then the screen return to the black standby screen. When the participant gives an ambiguous answer, the monitor remains in the black standby screen.

5.2.3 Study Procedure

An experimenter greeted the participant and invited him/her to the desk to complete the consent form and pre-survey measures. The experimenter gave the participant ample time to read through the consent form, then provided two more pre-survey forms to complete (demographic information and predispositions).

After filling out all forms, the experiment started by explaining the study procedures to the participant. To avoid any potential biases or differences, the experimenter explained the study procedures according to the pre-approved script; the experimenter first explained the medication-sorting task and then asked to run the trial test. If the participant did not have questions, the experimenter explained the 3-back auditory tasks and also ran the trial tasks. The trial could be performed as many times as the participant wanted.

Once the participant understood these two tasks, the experimenter explained the motor-cognition dual task and informed the participant that he/she needed to run these two tasks simultaneously. After completing the task introduction and trials, the compensation guideline was also informed to the participant.

After the participant's questions were answered, the study was started. Each participant's between-subject group was predetermined before he/she came to the

experiment. According to this decision, the participant was assigned to either the true or deception group. However, the participant was not informed of this group.

The order of within-subject conditions was also predetermined to maintain the counterbalance. Based on the order of within-subject conditions, the first feedback method (either robot or monitor) was setup and the study was started. After completing the first test set, the participant was asked to fill out two forms that measured his/her impressions and feelings of the feedback method and task load. While the participant answered the surveys, the experimenter changed the feedback method to the second within-subject condition and prepared the next set of study tasks. The second study set was conducted, and after, the same survey forms were asked to be filled out.

when the participant completed both sets of tests and forms, the experimenter informed the participant of the finish of the study and was asked to fill out the post-survey form, gathering his/her opinions on the ethics of robot deception. When all forms were complete, the experimenter debriefed the participant about the study and the concealment within the study. In this step, the purpose of this study was revealed and the participant was informed of his/her between-subject group. The experimenter answered any questions that the participant had, and was asked to sign the debriefing form to get the participant's approval to use the study data. In this debriefing session, the participant was also informed that the compensation guideline was just an experimental method, and that the participant would receive the full amount of compensation regardless of their performance. Finally, the experimenter provided the compensation and concluded the study.

5.2.4 Measurements

To answer the research hypothesis, multiple objective and subjective measures are gathered from the study as dependent variables. The main purpose of this study

is to observe the benefits of a robot's deceptive feedback. The benefits of robot deception in this HRI study can be measured in multiple ways. First, by observing the performance changes, the effects of deceptive feedback can be determined. For this purpose, objective measures of human's performance are collected. For example, if correctness of tasks is increase or decreased, it can indicate the changes of human's performance. In addition, participants' emotional status can also reflect the effects of robot deception. In other words, if a human's frustration level decreases or his/her motivation increases, we can analyze those changes as beneficial effects.

To measure the level of subjects' frustration and other emotions, multiple self-reported measures are gathered. Participants are asked to perform two sessions according to the two within-subject conditions; robot and monitor feedback. After each session, participants are asked to fill out the survey form, which asks the participants their impressions of each feedback method (Appendix B.4).Similarily, subjective measures such as participants' impressions of workloads are also gathered using NASA's Task Load Index (TLX) [67]. NASA's TLX consists of six questions to measure people/s workloads in different aspects. The six questions are about mental demand, physical demand, temporal demand, performance demand, effort, and frustration. Participants can answer those six questions in 21 gradations from 'very low' to 'very high.' Figure 39 shows this survey.

Finally, because robot deception is an ethically sensitive topic, participants' ethical opinions are also collected at the end of the study. The study results and more details will be explained in the robot ethics chapter later (Chapter 6).

Objective Measures

- Correctness of 3-back task: measuring the number of correct answers out of entire ten 3-back questions

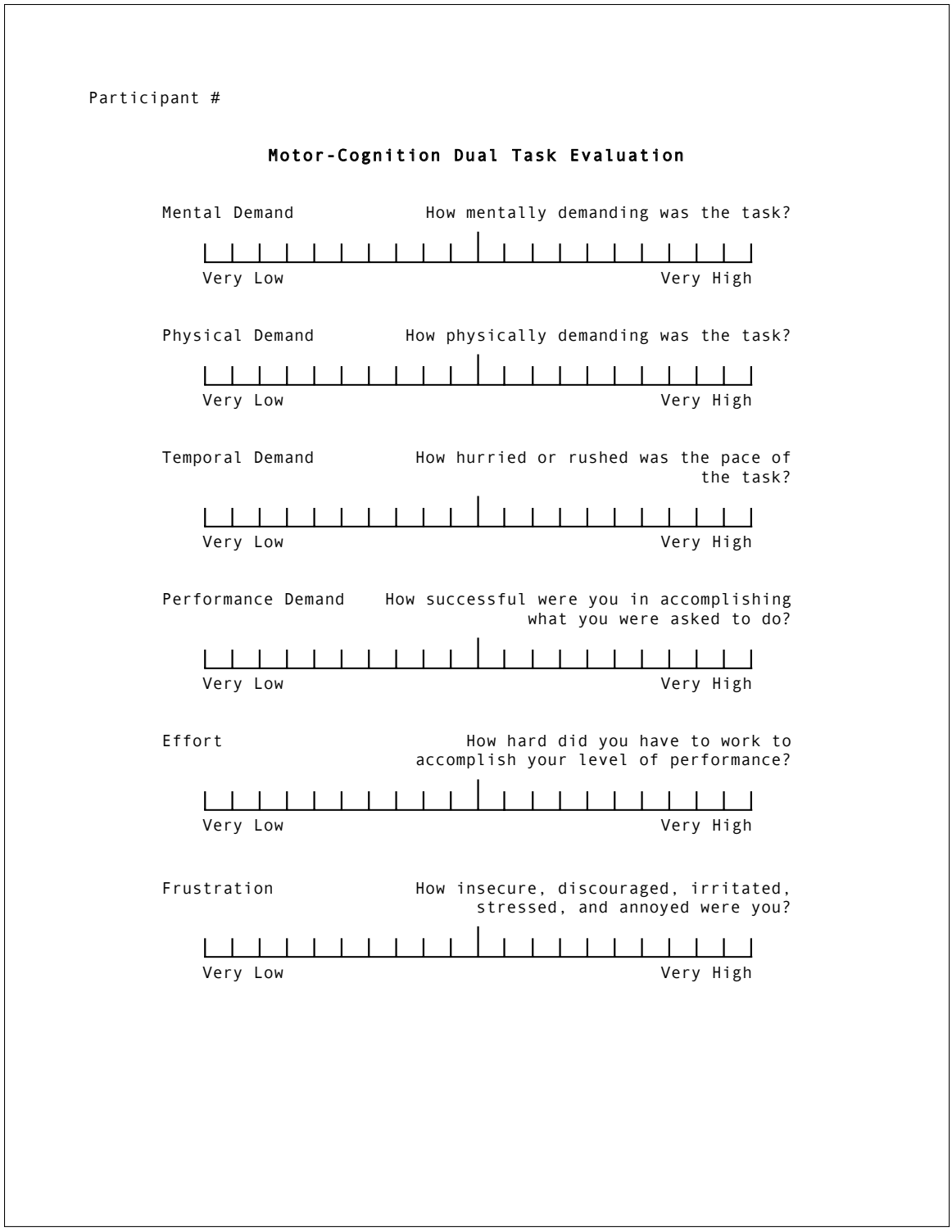


Figure 39: Self-reported measure: NASA's TLX is collected after each task session.

- Correctness of medication-sorting task: measuring the number of correctly-sorted results out of the entire eight medication-sorting tasks
- Time to complete the entire task (seconds): measuring the time from pushing the start button of the instruction (in iPad) to pushing the finish button
- Time to answer each n-back task (seconds): measuring the time from the end of n-back task question to the start of participant's answer
- Time to complete each individual medication sorting task (seconds): logging the time from the time to start the current task page to pushing the finish button

Subjective Measures

- Impressions of the robot feedback and the monitor feedback
- Impressions of the task levels (NASA's Task Load Index [67])
- Impressions/Opinions of robot deception

5.3 Study Results and Discussions

5.3.1 Demographic Information

A total of 34 subjects were recruited (22 females and 12 males). Since the task in the study was designed based on elderly people's rehabilitation tasks, the older adult population (over 55 years old) were recruited. Participants were recruited using flyers, email messages, as well as through word of mouth (Appendix B). The email was sent to mailing lists that are tied to an elderly group (Georgia Tech's Silver Jacket mailing list). Georgia Tech's Human Factors and Aging Lab also provided a list of potential older adults participants and their contact information according to the IRB approval, and this pool was also used for recruiting participants.

Table 20: Demographic Information from 34 HRI study participants

The highest level of education		
	# of Participants	**Percentage**
High School	2	5.8%
Bachelor's	16	47%
Master's	8	23.5%
PhD's	1	2.9%
other	7	20.5%
Technical level		
Yes	8	23.52%
Somewhat	10	29.41%
No	16	47.06%
Computer experience		
None	0	0%
Limited	2	5.88%
User Level	13	38.24%
Advanced User	16	47.06%
Programmer Level	2	5.88%
Advanced Programmer	1	2.94%
Prior experience with robots		
Never	30	88.24%
Very limited interaction	3	8.82%
Interaction experience with military robots	0	0%
Interaction experience with industrial robots	0	0%
Interaction experience with entertainment or educational robots	0	0%
Interaction experience with humanoid robots	1	2.94%

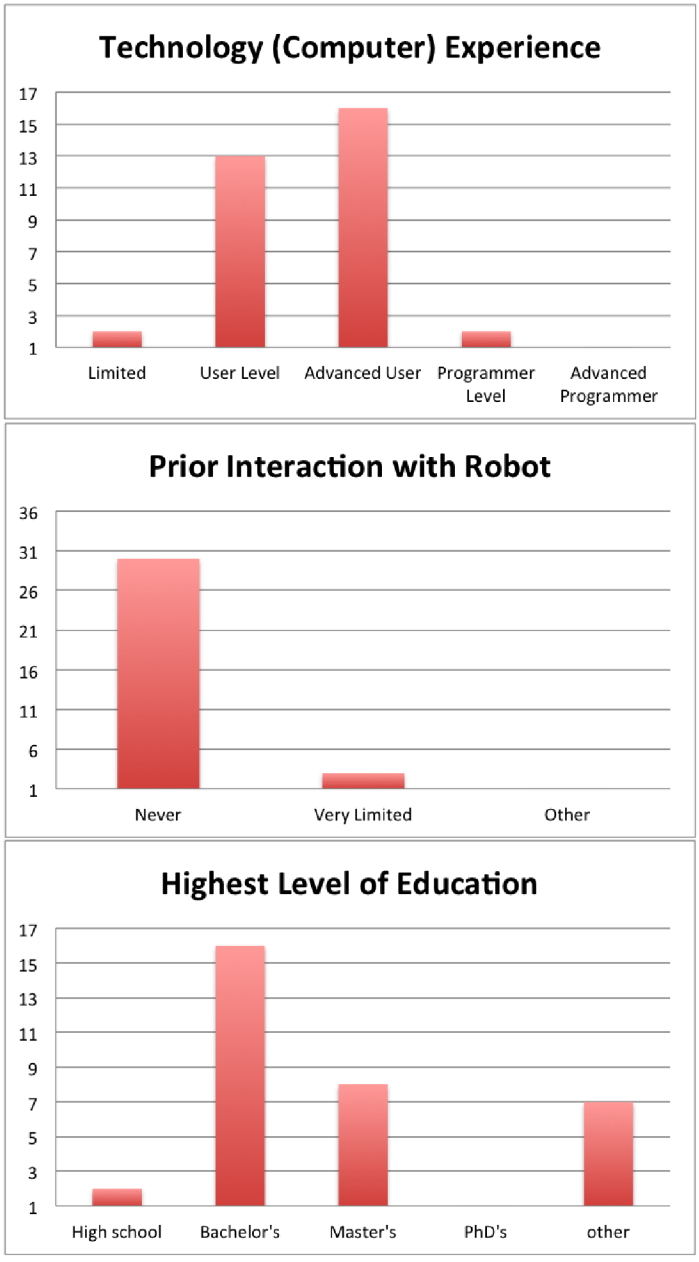


Figure 40: Demographic Information Chart: Technology level, Robot interaction experience, and Education level

The average age of the subjects is 69.12 years old (std:8.17, min: 58, max: 95). The basic demographic information for all subjects is shown in Table 20 and Figure 40. As illustrated, most participants (88.24%) did not have any prior experience with robots. Since participants were asked to use an iPad during the task, their technical level was asked and the results revealed that the average computer experience and technical level are good enough to run the study.

Subjects are assigned to one of the between-subject groups in the study. Feelings/emotions that people can have while interacting with a robot can be variable across different people; therefore, even if the experimental environment is the same, it is difficult to say the study result is normalized unless those groups have a composition such that the average predisposition is roughly the same. This can be ensured by finding no significant difference between the two between-subject groups on predisposition and personal trait measures taken by each participant. For this purpose, the Negative Attitudes towards Robots Scale (NARS) data have been gathered from the subjects via pre-survey [109]. This survey asks the subjects about their impressions and attitudes to robots in general, and as a result, it enables researchers to understand whether one group between conditions has disproportionately more people who are uncomfortable with social robots. The survey form is attached in Appendix B.3. When comparing the NARS survey results between the true and deception groups, the t-test revealed no significant differences ($p\text{-}value = 0.32 > 0.05$); therefore, the study can claim validity when comparing other measures between these two groups (true and deception groups) to support my research hypothesis.

5.3.2 Effects of robot deception

The main research question that this HRI study aims to answer is whether a robot's other-oriented deception can truly benefit human subjects.

Research Hypothesis 1: A robot's deceptive feedback (reaction) can positively affect a human's performance in the task.

First, it is observed how the subjects performed the 3-back auditory task questions in true and deception conditions. To figure out the performance benefit, objective measures were analyzed. The number of questions the subjects answered correctly or incorrectly are observed and analyzed. To test the effects of other-oriented robot deception, the data between true and deception conditions in the robot feedback group is compared. In the deception condition, the cases where a robot deceptively showed positive feedback to subjects' incorrect answers were counted as an incorrect answer. As shown in Table 21, in the true condition, 6.6 correct answers and 3.4 incorrect answers were observed on average ($\sigma^2 = 1.95$). In the deception condition, subjects answered the questions correctly 5.33 times and incorrectly 4.66 times on average ($\sigma^2 = 0.97$). The average number of times that the robot provided deceptive feedback is 1.93 (σ^2: 0.703, min: 1, max: 3). However, the t-test revealed no significant differences for this objective measure between true and deception conditions (p-*value* $= 0.5 > 0.05$), and therefore, the benefits of other-oriented robot deception were not observed in terms of a human's performance in this task.

Research Hypothesis 2: A robot's deceptive feedback (reaction) can reduce a human's frustration level in the task.

The deceived humans can also receive emotional benefits from robot deception. To test this research hypothesis, several self-report measures were collected from the

Table 21: Task performance: Average number of correct and incorrect answers from the true and deception condition

	True Condition	**Deception Condition**
The average number of correct answers	6.6 ($\sigma^2 = 1.95$)	4.66 ($\sigma^2 = 0.97$)

Table 22: NASA's TLX results: Average ratings from Deception and True groups; Scale: 0 (very low) - 21 (very high)

Robot Feedback Condition			
TLX Question	**Deception Group**	**True Group**	**p-value**
Mental Demand	10.647	12.823	0.174
Physical Demand	4.529	6.352	0.135
Temporal Demand	9.058	13.235	0.009
Performance Demand	11.647	12.117	0.384
Effort	10.47	14.411	0.006
Frustration	6.47	9.588	0.044
Monitor Feedback Condition			
TLX Question	**Deception Group**	**True Group**	**p-value**
Mental Demand	11.176	14.764	0.1945
Physical Demand	4.647	5.588	0.175
Temporal Demand	9.411	13.47	0.137
Performance Demand	12	9.529	0.0721
Effort	12.764	13.47	0.3
Frustration	7.411	10.058	0.119

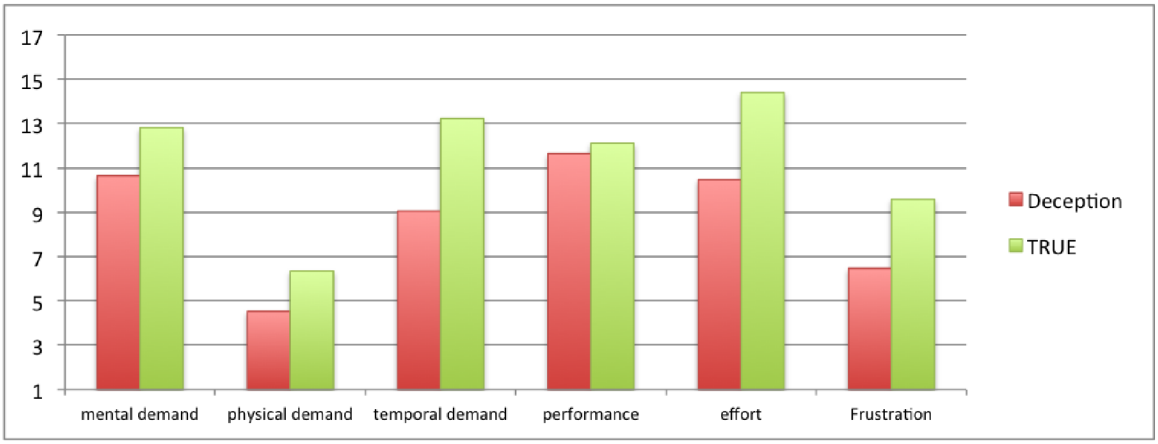


Figure 41: NASA's TLX results from Robot feedback condition: Red-average ratings from Deception group, Green-average ratings from True group

subjects, and some of the results illustrate interesting findings. To measure subjects' workload and frustration level, a NASA Task Load Index [67] was collected right after each task set. NASA's TLX questionnaires ask the subjects to rate six questions in 21 gradations on the scales (0-very low to 21-very high). The six questions are about mental demand, physical demand, temporal demand, performance demand, effort, and frustration. As shown in Table 22 and Figure 41, TLX ratings for all six questions are greater in true condition compared to deception condition. In particular, significant differences are observed between true and deception conditions in the following three of the six questions.

1) Frustration: How insecure, discouraged, irritated, stressed, and annoyed were you? (Two-sampled t-test's $p\text{-}value = 0.044 < 0.05$)

2) Temporal demand: How hurried or rushed was the pace of the task?
(Two-sampled t- test's $p\text{-}value = 0.009 < 0.05$)

3) Effort: How hard did you work to accomplish your level of performance?
(Two sampled t- test's $p\text{-}value = 0.006 < 0.05$)

As shown above, the answers to the three questions above showed significant differences between true and deception conditions. In particular, as illustrated in Table 22, an average rating of the deception group's frustration question was measured as 6.47, which is significantly lower than the true group's average ratings (9.58). In 21 gradations of the scales, 0 means "very low" and 21 indicates "very high." Therefore, low ratings can be interpreted as a lower frustration level: thus, the result can claim that a robot's deceptive feedback can significantly reduce subjects' frustration level for this task. Similar to this analysis, answers in temporal demand and effort questions also show significant differences between two groups and the results can be interpreted as that the subjects felt the task required relatively lower times (deception group: 9.057 vs. true group: 13.235) and effort (deception group: 10.47 vs. true group:

Table 23: Task performance: Average number of correct and incorrect answers from the robot and monitor feedback condition

	Robot feedback	**Monitor feedback**
The average number of correct answers	6.41 ($\sigma^2 = 1.989$)	5.38 ($\sigma^2 = 1.96$)

14.411) in the deception condition. This may result as the deceived humans are motivated to engage the task more and achieve the task quickly. In sum, the results can affirm that ***a robot's deceptive feedback positively affected a human's frustration level, according to the self-report measures***.

5.3.3 Effects of a robot's embodiment

Research Hypothesis: A physical robot's deceptive feedback can increase a human being's engagement and enjoyment in the performance task when compared to non-robotic feedback.

Another hypothesis is that the human-like robot's embodiment could help the elderly to engage in tasks and lead to a more enjoyable rehabilitation experience. For this purpose, participants were asked to perform the task set twice with two different within-subject conditions; monitor feedback and robot feedback. As shown in Table 23, in the robot feedback condition, the average number of correct answers is slightly but not significantly greater than in the monitor feedback condition ($p\text{-}value = 0.51 > 0.05$).

However, several self-reported measures showed significant differences. The responses are on a five-point Likert-scale and the ranges of ratings are different for each question where definitions of rating 1 and rating 5 are opposite of each other (Appendix B.4). As shown in Table 24 and Figure 42, subjects were impressed that ***the robot feedback was significantly more noticeable, helpful, trustful, and interactive than the monitor feedback***.

There were several interesting comments from subjects, which can reflect that

Table 24: Self-reported measures: Impressions of a robot or monitor feedback; Average ratings (standard deviation) and p-value

Question: During this task, feedback from the (robot/monitor screen) was:			
Scales	**Robot**	**Monitor**	**p-value**
Noticeable(1) - Ignorable(5)	2.44 (1.3)	3 (1.477)	0.023
Interfering(1) - Minding its own business(5)	3.52 (0.99)	3.647 (1.04)	0.29
Annoying(1) - Inoffensive(5)	2.82 (1.19)	3.911 (1.16)	0.383
Irritating(1) - Undemanding(5)	3.94 (1.2)	3.97 (1.08)	0.46
Bothersome(1) - Quiet(5)	3.5 (0.96)	3.82 (1.19)	0.124
Question: In your opinion, (robot/monitor screen) appeared:			
Scales	**Robot**	**Monitor**	**p-value**
Fake(1) - Natural(5)	3.44 (1.3)	3.79 (1.22)	0.105
Machinelike(1) - Humanlike(5)	2.64 (1.15)	1.97 (0.93)	0.0006
Unconscious(1) - Conscious(5)	4.08 (1.02)	2.94 (1.07)	2.59E-08
Artificial(1) - Lifelike(5)	3.17 (1.24)	1.91(0.96)	2.62E-08
Inert(1) - Interactive(5)	3.94 (1.09)	2.52 (1.46)	3.49E-05
Question: During the task, feedback from (robot/monitor screen) was:			
Scales	**Robot**	**Monitor**	**p-value**
Unhelpful(1) - Helpful(5)	3.52 (1.21)	3.14 (1.32)	0.0367
Not Trustful(1) - Trustful(5)	4.41 (0.74)	4.02 (1.05)	0.045
Boring(1) - Enjoyable(5)	4.26 (0.96)	3.22 (1.3)	0.044

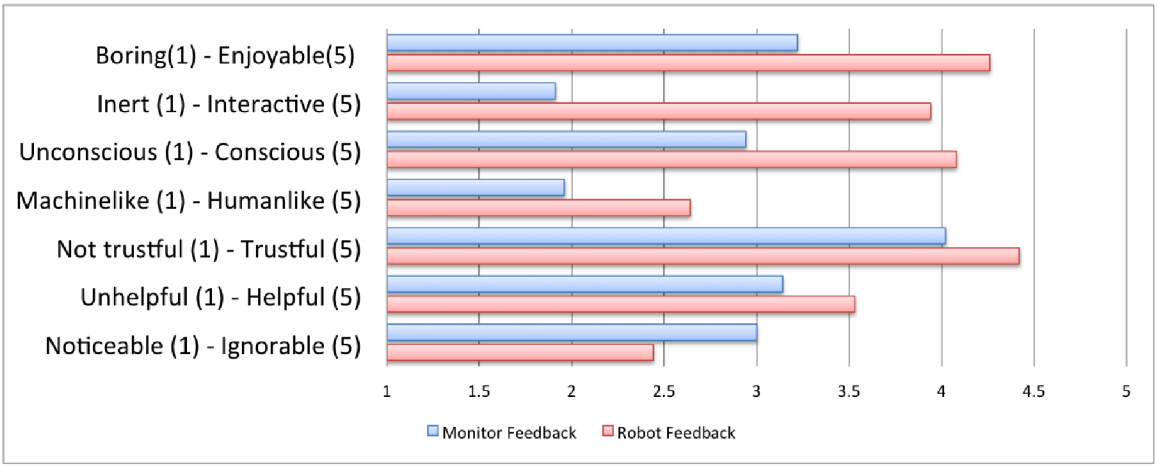


Figure 42: Post-survey results: Impressions of robot (red) or monitor (blue) feedback during the task

subjects were more enjoyed and received positive effects with robot feedback.

> "Robot feedback: more enjoyable to do the task"
>
> "There was a sense of wanting to please the robot, which was not there with the computer monitor."
>
> "This was a lot of fun. I enjoyed interacting with the robot very much. he was very cute."
>
> "Great interaction with Nao, just so enjoyed!"

The results reflect that subjects had a more enjoyable rehabilitation experience with robot feedback and robot feedback worked as a positive reinforcement for participants to engage more in the task.

5.3.4 Ethical Implications of other-oriented robot deception

Research Hypothesis: Robot deception is acceptable if it is used exclusively for the deceived human's benefit and advantage.

It is essential to discuss the ethical aspects of robot deception. Self-reported survey measures were also gathered to access subjects' opinions on the use of robot deception

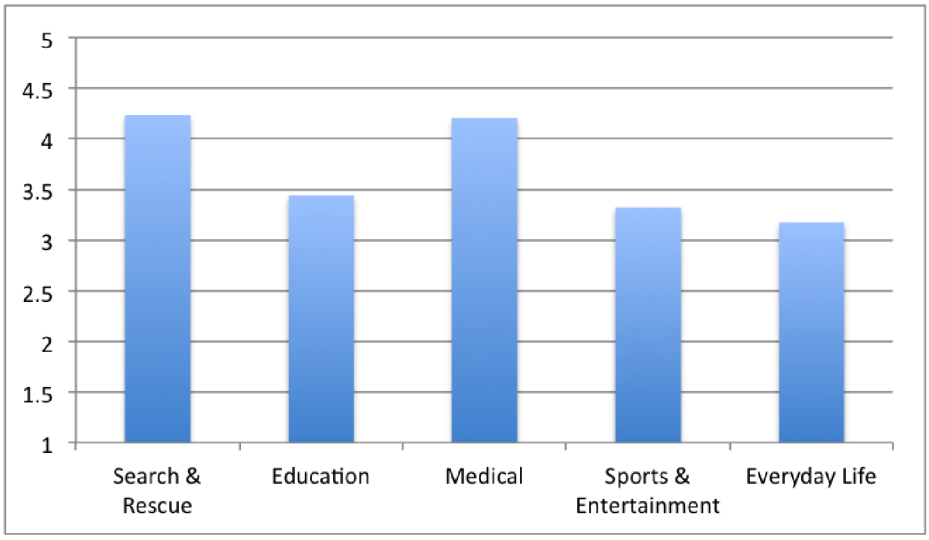


Figure 43: Post-survey results; Ethical Question: I can accept robot deception in [Context in x-axis] if it is strictly used only to benefit humans; Scales in y-axis: 1-strongly disagree, 5-strongly agree

at the end of HRI study. The survey made several ethical statements and the response was a rating on a five-point Likert scale (the ratings ranged from 1-strongly disagree to 5-strongly agree). Questions asked broadly whether they would accept a robot's other-oriented deception. According to the results, regarding the statement: "A robot can hide/misrepresent information if it can help humans," the average answer was 3.24 ($\sigma^2 = 0.88$). In addition, the statement: "The robot should always be honest in any circumstance," received on average an answer of 3.0 ($\sigma^2 = 1.12$). "Robot can intentionally/unintentionally deceive humans if it's in an appropriate situation" was rated 3.38 ($\sigma^2 = 1.18$) on average. In other words, these average ratings are around 3 points (undecided), which means the results illustrate that we cannot determine the ethical acceptability of robot deception with these broad and high-level statements.

However, when specified the situation (context) are provided, subjects' acceptance rates slightly increased as shown in Figure 43. Here, survey questions asked the statement: "I can accept robot deception in [certain context] if it is strictly used only to benefit humans" using five different contexts as shown in Figure 43. The results can form an ethical implication of robot deception such as "People can accept the use

of other-oriented robot deception when an appropriate and specific context is clearly determined." In sum, ***the strong motives of deception in each context should be discussed and validated when other-oriented robot deception is used in HRI context***.

The results revealed an interesting ethical implication such that motives and appropriate contexts should clearly be declared before applying other-oriented deception to HRI situations. Even though such an interesting ethical implication related to other-oriented robot deception was found, this result has limitations since survey responses were collected in a small number of people and also they are in a target population (aging group). To generalize my argumentation, it is necessary to gather more survey responses from the general public. For this purpose, a follow-up web-based survey was conducted and the results are in Chapter 6.

5.3.5 Summary

With the increasing use of social robots in HRI, deception can be an important capability similar to its use by humans. In particular, I assert that robot deception should be used when it offers strong motives to benefit the deceived humans in an appropriate HRI context. In this HRI study, I presented an HRI context that potentially contains motives for a robot's other-oriented deception: elderly persons' rehabilitation tasks and Parkinson's patients' daily activities. By conducting an HRI study in this context with 34 older adults, I validated several research hypotheses. First, the results reveal that a robot's deceptive feedback can positively affect participants' frustration level. Therefore, it supports the research hypothesis "a robot's deceptive feedback can positively affect a human's frustration in the task." In addition, to argue the effects of a deceptive robot's embodiment, I also compared the results between robot feedback and monitor feedback. From the results, it can result that people can have more enjoyable rehabilitation experience with robot feedback. In sum, this

HRI study concludes that a robot's deceptive feedback has the potential to help the deceived subjects' engagement and enjoyment in the task(s).

5.4 Extended HRI study

As described in the previous sections, the HRI study results revealed the beneficial effects of other-oriented robot deception in elderly's rehabilitation tasks. The robot's deceptive behaviors in the previous HRI study were based on a deceptive action generation model, which has been illustrated in section 4.2. However, deceptive action selection (when to perform the deceptive or true feedback) was pre-defined from the pilot study results. In other words, a robot deterministically performed the deceptive feedback only when participants answered the question incorrectly more than twice. To validate the dynamic deceptive action selection model for a robot, an extended HRI study has been conducted.

The other-oriented deceptive action selection model was proposed in the computational model in section 4. In this extension, the proposed action selection model was developed and programmed into the robot. After implementation, the same procedures as the previous HRI study were performed with the same target population again to validate the model.

Action selection model using CBR

As described in section 4.3, the deceptive action selection model in this research has been developed using a case-based reasoning mechanism. Multiple factors need to be predetermined and implemented to apply to the deceptive action selection model. First, case $c = [s, a, r]$ of the CBR model should be defined. As described in section 4.3, a case c consists of state (s), action (a), and benefit (r). State features, which describe the current status of the participants, have been empirically determined as $s =< f_{current}, f_{shortterm}, f_{longterm}, f_{timetoanswer} >$ by observing the previous HRI study

results. These features demonstrate the subject's current and previous performance, and the definitions of each state are as follows.

$f_{current} = \{x | x \in \{1, 0\}\}$: correctness in the current question where

$$x = \begin{cases} 1 & \text{if the participant answers the current question correctly} \\ 0 & \text{Otherwise} \end{cases} \tag{9}$$

$f_{shortterm} = \{x | x \in \{1, 0\}\}$: correctness in the previous question where

$$x = \begin{cases} 1 & \text{if the participant answers the previous question correctly} \\ 0 & \text{Otherwise} \end{cases} \tag{10}$$

$f_{longterm} = \{x | 0 \leq x \leq 10, x \in \mathbb{Z}\}$: correctness rate from the first to the current question where

$$x = \lfloor 10 \cdot \frac{Numberofcorrectanswers}{Numberofquestions} \rfloor \tag{11}$$

$f_{timetoanswer} = \{x | x \in \{short, average, long\}\}$: Time to answer the question where

$$x = \begin{cases} short & \text{if } t < 2 \ seconds \\ average & \text{if } 2 \ seconds \leq t \leq 3 \ seconds \\ long & \text{if } t > 3 \ seconds \end{cases} \tag{12}$$

Here, t is the time from right after the end of the question to the beginning of the participant's answer. The conditional times for short, average, and long categories are determined empirically by observing the previous HRI study results. According to the previous study, the average time to answer the question was 2.35 seconds (min=

1, max= 5.3). Therefore, the time to determine short and long periods are determined as two and three seconds in this model.

The action is the same as the previous HRI study. There are positive and negative feedback gestures. True and deceptive responses can be determined based on the participant's current answer.

$A = \{a_{truepositive}, a_{truenegative}, a_{deceptivepositive}, a_{deceptivenegative}\}$

If $f_{current} == 1$, then action can be $a_{truepositive}$ or $a_{deceptivenegative}$.

If $f_{current} == 0$, then action can be $a_{truenegative}$ or $a_{deceptivepositive}$.

Finally, benefit is an essential feature in the deceptive action selection model, because a robot should select true or deceptive feedback action to maximize the participants' benefits. The benefits are determined by whether performance improves, worsens, or does not change. By comparing the correctness on the previous and the current question, the benefit can be determined as shown below.

$$R = \{r | r \in R, r = r_{changes} * 0.6 + f_{next} * 0.4\} \text{ where}$$

$$r_{changes} = \begin{cases} -1 & \text{if } f_{current} == 1 \text{ and } f_{next} == 0 \\ 1 & \text{if } f_{current} == 0 \text{ and } f_{next} == 1 \\ 0 & \text{Otherwise; no changes} \end{cases} \tag{13}$$

$$f_{next} = \begin{cases} 1 & \text{if participant's answer from the next question is correct} \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

Two parameters are used to determine the benefits of the subjects. First, the participant's performance in the next question is used to observe whether a robot's

feedback positively or negatively affects the participant. For this, f_{next} parameter is used to calculate the benefit r, and it can be determined as shown in equation 14. More importantly, improvement of the participant's performance can be an essential feature to indicate the effects of feedback. Here, $r_{changes}$ indicates whether the participant's performance changes compared to the previous performance. To measure this, the participant's correctness to the current and the next questions is observed. As shown in equation 13, when the participant answers the current question incorrectly, but answers the next question correctly, this can be interpreted as increased participant's performance and received "benefits" from a robot's feedback. Parameter $r_{changes}$ is defined to observe such performance changes from the correctness of current question to the next question. By observing this change, we can determine whether a robot's true/deceptive feedback affects human's performance, as described in equation 13. The weights for these two features are empirically determined by the experimenter. In the current version, the experimenter observed the previous HRI study and manually determined the weights as 0.6 for parameter $r_{changes}$ and 0.4 for parameter f_{next}. Note that, as shown in this equation, a robot should observe the correctness for the next question to calculate the benefit. Therefore, the benefit is determined in the next question phase by observing f_{next}. As a result, the casebase update is also determined one step later.

In addition to the case definition, the initial casebase should be also populated. The experimenter can determine the initial casebase by analyzing the experimental results in the previous HRI study. The predetermined initial casebase in this study has been empirically determined as shown in Table 25.

Based on these definitions, a robot's other-oriented action selection model has been implemented and applied to the robotic system. The reference source code is attached in appendix (Deceptive Action Selection main class in Appendix D.1, Casebase class in Appendix D.2).

Table 25: Extended HRI study Initial Casebase (predetermined)

Case#	**State S**				**Action**	**Benefit**
	$f_{current}$	$f_{shortterm}$	$f_{longterm}$	$f_{timetoanswer}$		
1	1	1	10	short	$a_{truepositive}$	0.9
2	0	0	5	long	$a_{deceptivepositive}$	0.7
3	1	0	5	average	$a_{truepositive}$	0.7
4	0	0	3	long	$a_{deceptivepositive}$	0.7
5	0	0	2	short	$a_{deceptivepositive}$	0.3
6	1	0	2	short	$a_{deceptivenegative}$	0.3
7	0	1	9	average	$a_{truenegative}$	0.5

Study Results and Discussions

The main research hypothesis related to other-oriented robot deception has already been proven from the results of the previous HRI study, and the goal of this extended HRI study is to observe whether computational models of other-oriented deception work properly. Since the computational model's proper working is only needed to be evaluated, we included only five more participants. Participants were recruited and the same HRI study was run. Again, only elderly people were recruited and the HRI study procedures remained the same. Table 26 illustrates the basic demographic information.

The only difference from the previous study was how the robot determined the moment to perform deceptive feedback. In the previous study described above, a robot performed deceptive feedback when participants incorrectly answer the questions more than twice. In contrast to the previous HRI study, in this extended study a robot can dynamically determine the moment to perform deceptive feedback by observing the previous experiences (casebase).

The only difference from the previous study was how the robot determined the moment to perform deceptive feedback. In the previous study described above, a

Table 26: Demographic Information from Five extended HRI study participants

Average Age: 67.2 (min: 57, max: 73, σ^2: 6.18)		
The highest level of education		
	# of Participants	**Percentage**
High School	0	0%
Bachelor's	2	40%
Master's	1	20%
PhD's	1	20%
other	1	20%
Technical level		
Yes	2	40%
Somewhat	2	40%
No	1	20%
Computer experience		
None	0	0%
Limited	0	0%
User Level	1	20%
Advanced User	4	80%
Programmer Level	0	0%
Advanced Programmer	0	0%
Prior experience with robots		
Never	5	100%
Very limited interaction	0	0%
Interaction experience with military robots	0	0%
Interaction experience with industrial robots	0	0%
Interaction experience with entertainment or educational robots	0	0%
Interaction experience with humanoid robots	0	0%

Table 27: Extended HRI study results from all five participant: task performance (number of correct answers) and deceptive action generation results (number of deceptive feedback)

Participants	number of correct answers	number of deception	number of $a_{deceptivenegative}$	number of $a_{deceptivepositive}$
1	6	4	1	3
2	7	4	2	2
3	4	5	2	3
4	8	1	0	1
5	6	4	1	3
average	6.2	3.6	1.2	2.4
σ^2	1.483	1.5166	0.837	0.894

robot performed deceptive feedback when participants incorrectly answer the questions more than twice. Different from the previous HRI study, in this extended study, a robot can dynamically determine the moment to perform deceptive feedback by observing the previous experienceeriences (casebase). Therefore, it is expected that a robot generates deceptive behaviors more or less than the deterministic models according to prior experiences. Table 27 illustrates the results of this extended HRI study. As shown in the table, it is observed that the average number of deceptions increased to 3.6 (σ^2: 1.5166) compared to the average number of deceptions using the deterministic model in the previous HRI study (1.93, σ^2: 0.703). From this observation, it can be argued that this study reveals this computational model could make a robot perform deceptive feedback more dynamically in this study domain.

As illustrated in the robot's action definition, there are two ways for the robot to perform the deceptive feedback. First, a robot can deceptively perform positive gesture when participants answered incorrectly ($a_{deceptivepositive}$). In addition, a robot can also perform deceptive feedback by showing negative gesture even though participants answer the question correctly ($a_{deceptivenegative}$). In the previous study, the robot

only performed deceptive "positive" feedback ($a_{deceptiveposition}$) when participants answer the question incorrectly. However, since the robot in this action selection model can dynamically determine the deceptive feedback based on experience, the robot can also perform deceptive "negative" feedback ($a_{deceptivenegative}$).

According to the results of the study, deceptive "positive" feedback was observed an average of 2.1 times. Additionally, deceptive "negative" feedback was performed by the robot an average of 1.2 times. Due to the small study participants in the extended study, a direct comparison of the previous HRI study results to the current results is inadequate. However, it can be argued that these results are still interesting, since both deceptive "positive" feedback and deceptive "negative" feedback were observed to be performed from the robot and this can be extended to argue that the dynamic action selection model could potentially work with real human subjects.

In the previous HRI study, the deceived human's benefit to robot deception were verified by frustration level and positive impressions to robot feedback. In this extended study, NASA's TLX and post surveys were also collected to measure these factors, and Tables 29 and 28 illustrate the results of these self-reported measures. Again, due to the small number of participants in this extended study, it is inadequate to claim that the participants' frustration levels are similarly low to in comparison to the levels in the previous HRI study. However, the average rating 4.8 ($\sigma^2 = 3.898$) is a significantly low number in comparison to the median of the scale. Even though it is inadequate to claim participants' benefits statistically, these results can be an indication of the potential benefits of other-oriented robot deception with this dynamic deceptive action selection model.

5.5 Summary

This dissertation demonstrates whether a robot's other-oriented deception can be applicable in an appropriate HRI context and whether it can truly benefit the deceived

Table 28: Extended HRI study self-reported measures: Impressions to a deceptive robot, average ratings from the five participants

Question: In your opinion, during this task, feedback from the robot was:	
Scales(1-5)	Average Rating
Noticeable(1) - Ignorable(5)	4.4 ($\sigma^2 = 0.547$)
Interfering(1) - Minding its own business(5)	3.8 ($\sigma^2 = 0.836$)
Annoying(1) - Inoffensive(5)	3.8 ($\sigma^2 = 1.303$)
Irritating(1) - Undemanding(5)	4 ($\sigma^2 = 1.0$)
Bothersome(1) - Quiet(5)	3.4 ($\sigma^2 = 0.547$)
Question: In your opinion, the robot appeared:	
Scales(1-5)	Average Rating
Fake(1) - Natural(5)	2.8 ($\sigma^2 = 1.308$)
Machinelike(1) - Humanlike(5)	2 ($\sigma^2 = 0.707$)
Unconscious(1) - Conscious(5)	4 ($\sigma^2 = 1.0$)
Artificial(1) - Lifelike(5)	3 ($\sigma^2 = 1.224$)
Inert(1) - Interactive(5)	3.6 ($\sigma^2 = 1.14$)
Question: During the task, feedback from the robot was:	
Scales(1-5)	Average Rating
Unhelpful(1) - Helpful(5)	3.4 ($\sigma^2 = 1.14$)
Not Trustful(1) - Trustful(5)	3.8 ($\sigma^2 = 1.303$)
Boring(1) - Enjoyable(5)	4.8 ($\sigma^2 = 0.447$)

Table 29: NASA's TLX results: Average ratings from Extended HRI study's participants; Scale: 0 (very low) - 21 (very high)

TLX Question	**Average Rating**
Mental Demand	11.2 ($\sigma^2 = 4.02$)
Physical Demand	5.2 ($\sigma^2 = 5.16$)
Temporal Demand	9.6 ($\sigma^2 = 4.92$)
Performance Demand	14 ($\sigma^2 = 5.09$)
Effort	13.6 ($\sigma^2 = 4.15$)
Frustration	4.8 ($\sigma^2 = 3.898$)

human partners. Based on the preliminary research and development of a computational model, this chapter discusses whether and how the research hypotheses in this dissertation can be supported via several HRI studies and their results.

The main goal of the HRI study was to demonstrate that a robot's other-oriented deception can benefit humans in an appropriate situation. When applying other-oriented deception to a robotic system, it is essential to first validate whether or not a target context contains sufficient motive to perform other-oriented deception. This can be determined by observing deceptions in human-human cases. In other words, if humans perform benevolent deception in some specific contexts, we can argue that those contexts potentially have motives for other-oriented robot deception. For this purpose, multiple situations in which humans frequently use other-oriented deception were reviewed. Finally, among different contexts, elderly's rehabilitation was selected as an HRI study context in this research.

The design of the HRI study was inspired by elderly people's rehabilitation. In particular, the elderly's motor-cognition dual task was selected as the HRI study task. When the participants performed the assigned motor-cognition dual tasks, the robot assistant was placed next to them and provided feedback on their performance. Basically, the robot provided positive feedback when participants correctly solved the tasks and negative feedback when the tasks were solved incorrectly. However, the robot was also capable of performing other-oriented deceptive feedback, thus motivating and encouraging the participants. Occasionally, the robot also generates deceptive positive feedback even when the participants incorrectly answer the questions.

This HRI study aimed to observe whether a robot's other-oriented deceptive feedback could help human subjects increase their performance in rehabilitation tasks. 34 participants over 55 years old were recruited and a two by two mixed-subject study was conducted. The results revealed that a robot's other-oriented deception

can reduce the participant's frustration and increase the participant's enjoyment during the rehabilitation tasks. Finally, from these results, the following two research hypotheses were validated and the benefits of a robot's other-oriented deception were demonstrated.

- A robot's deceptive feedback (reaction) can positively affect the deceived humans in terms of frustration level in the performance task.
- A physical robot's deceptive feedback can increase a human being's engagement and enjoyment in the performance task.

As argued in this dissertation, a robot's other-oriented deception can positively affect the deceived humans in an appropriate situation. To further test my deceptive action selection model, an extended HRI study was also conducted. The HRI study procedures were the same, except the robot's action selection was made using the CBR-based deceptive action selection model. Five more participants were recruited to re-run the study. From the results, it is observed that the model enabled a robot to perform other-oriented deception dynamically.

The HRI study results in this chapter showed the successful use of other-oriented robot deception and revealed the potential benefits for the deceived humans in an appropriate context. Finally, robot deception is an ethically sensitive topic so participants' opinions were also collected from HRI studies. The results of these ethical questions are discussed in the following chapter.

CHAPTER VI

ROBOT DECEPTION AND ETHICAL ISSUES

The benefits of other-oriented robot deception have been discussed throughout this dissertation. First, computational models to demonstrate other-oriented deception were developed and proposed. Then, by conducting HRI studies, it was proven that other-oriented robot deception can potentially be used in an appropriate context. In addition, the proposed computational model was applied and evaluated. In sum, all of these research results revealed the beneficial use of other-oriented robot deception in HRI.

Despite the benefits of robot deception, relatively little research has been conducted to date on this subject, perhaps as a result of ethical considerations involving this somewhat controversial topic. Therefore, arguing ethical issues is essential in this research. To discuss ethical argumentation related to robot deception, literature reviews and survey research were conducted as described in the previous chapters. In particular, post-survey responses related to ethical questions were collected from the HRI study participants, as described in the previous chapters, to evaluate people's opinions on robot deception.

Multiple ethical implications were observed from the HRI study's results. However, the study has many limitations. Survey responses were collected from a small number of people (39 participants). In addition, the HRI study's design was focused on elderly people's rehabilitation, and so all of the participants were recruited from a specific target population (an aging group). Therefore, the findings from this study are difficult to generalize. To overcome these limitations, a web-based survey was

conducted to collect the opinions and impressions of robot deception from the general public. The ethical implications of robot deception, which were observed from the online survey's results, will be discussed in this chapter.

6.1 Robot Ethics

Robot ethics is a rapidly growing topic in many research areas including robotics, psychology, philosophy, and law [136, 51, 87, 152]. Researchers aim to discuss and understand the ethical consequences and implications when using robotic technology in human society. At the present time, a variety of different topics are argued in robot ethics, including eldercare and medical robots, autonomous robot missions in military situations, entertainment and service robots, and so on.

Initially, a main focus of robot ethics was the discussion of military robots [88, 153]. When robots are used in battlefields we face critical ethical questions. Is it ethically correct for a robot to autonomously decide to use lethal force?; in other words, can a robot kill the enemy? Who should take responsibility for any mistakes that might occur from the decision of an autonomous robot? Many other ethical arguments are currently being extensively discussed in robot ethics areas [111, 137].

By increasing the use of social robots, ethical arguments related to robots in human-robot interactions are also progressively expanding [87, 152, 143]. The use of service robots in healthcare contexts is one obvious area with many ethical arguments, as many service robots are developed to support the elderly or patients [106, 17]. In such cases, since autonomous robots can directly interact with those populations, people argue that the ethical consequences should be considered more carefully and ethical verifications should be required beforehand [166, 143, 145].

Most recently, many companies have announced that autonomous vehicles are ready to be deployed [177, 74]. Even though the technology for self-driving cars is sufficient, there are still many ethical dilemmas and questions. For example, the

classic ethical question, the Trolley problem [161], is now a practical issue in this area [74]. To resolve many ethical issues concerning autonomous vehicles, carmakers and researchers are developing algorithmic morality; however, this still requires extensive discussion for agreement.

Among different topics, robot deception is a more critical topic for argument. Practically, it is even complicated to state the ethical issues related to deception than in human cases. However, it is obvious that this issue should be carefully considered during the development and application of robot deception [10]. Many ethical questions can arise when deception is applied to the robotic system [87, 177]. For example, we can face ethical questions such as "Is deception acceptable even in humans?" or "Should a robot be allowed to lie?" Furthermore, since deception is related to trust [65], the discussion of deception is getting more important.

To discuss ethical issues of robot deception, it is necessary to review the fundamental moral theories of deception in philosophy. According to Kantian theory, deception or lies should always be prohibited, a standard outcome of any ethics classroom in the application of the Categorical Imperative [34]. By this standard, any deceptive behaviors or lies are morally incorrect, human or robot. The utilitarian perspective, on the other hand, argues that an action is morally right and acceptable if it leads to increasing total happiness over all relevant stakeholders [149]. By this perspective, I can also argue that if deception increases the total benefits among the involved relationships, it is ethically correct [149, 34]. More specifically, Bentham and Mill [148] argued that it is morally right if and only if any behaviors/acts produce overall increased happiness. In other words, an action is morally good if it provides overall benefits. This ethical theory is called act-utilitarianism.

Related to robot deception, Reynolds and Ishikawa [122] discussed the role of designers and robots and emphasized the importance of morally responsible entities. Arkin [13] also pointed how important it is in discussing the ethical justification of

robot deception.

In affective computing, researchers have argued that the use of emotional robots is deceptive. According to Coeckelbergh, emotional robots are deceptive since "1. Emotional robots intend to deceive with their "emotions." 2. Robotic emotions are unreal. 3. emotional robots pretend to be a kind of entity they are not. [35]" For example, when such robots are used in eldercare, peoples are led to believe that they are loved or cared for by the robots, and according to the definition, this can be a case of delusion [154]. Finally, by reformulating these three claims of emotional robot deception to ethical criteria, he argued that the situation can also be considered "ideal emotional communication" rather than deception.

More recently, Matthias suggested four criteria for robot deception [95]. He argued that by fulfilling four conditions, robot deception can be morally permissible. These four criteria are trust, autonomy, transparency, and safety. First, robot deception should not betray patients' trust by promoting patients' interests. Also, deception should support patients' autonomy by supporting them to make decisions and control the machines better. To be transparent to the patients, the fact that deception is happening should be suggested at some point in the conversational context. Finally and most importantly, deception should not lead to any harm to the patients.

There is also an HRI study related to the human moral stance and robot deception. In Kahn's HRI study [79], subjects were asked to play a game, and a humanoid robot guided and observed their performance. After completing the game, a robot debriefed the subjects but announced their achievement deceptively as being lower. Here, researchers observed that 65% of the participants expressed some level of moral accountability to the robot during the study. In addition, they also reported that by adding more interaction capabilities to the robot (in terms of speech communication skills), the rates to attribute moral accountability to the robot also increased. In sum, this study revealed that more socially-intelligent robot deception caused many

people to hold the robot morally accountable.

Researchers also sometimes argue the ethics of robot deception based on situations. Nijholt identified the potential situations or contexts in which artificial human partners will be deceptive (or not honest) by analyzing human-human cases [108]. He proposed four categories of situations in which natural deceptive interactions will be used or required in human-robot or human-computer interactions. These four categories are 1) conversations and dialogues; 2) commerce, negotiation, persuasion; 3) teaching, training, serious games; and 4) sports, games, and entertainment. There are also several ethical arguments about robot deception in different contexts, including economics and law [68], healthcare [95], and so on.

6.2 Ethical implications from the online survey's results

This dissertation aims to evaluate the benefits of a robot's deceptive behaviors in a situation involving human-robot interaction. To support my research hypothesis, human-subject studies were conducted and the results were analyzed in the previous chapter (chapter 5). In sum, the study evaluated whether or not a robot can deceive human subjects and assessed how those deceptive behaviors affect human subjects when performing a simple task. The results revealed that a robot's deceptive feedback can help human subjects better focus on a task involving both motor skills and cognition at the same time. Despite this result, robot deception remains an ethically sensitive topic. To better understand the ethical issues of robot deception, I also gathered the subjects' ethical opinions at the end of the HRI studies. According to the results, the participants were negative or undecided (about/less than 3.0 in average rating) on the concept of robot deception in general, even though it was described as other-oriented deception. For example, in the previous study, "A robot can hide/misrepresent information if it can help humans," was rated 3.24 ($\sigma^2 = 0.88$) where 1 is strongly disagree and 5 is strongly agree. However, when the situation (context) was

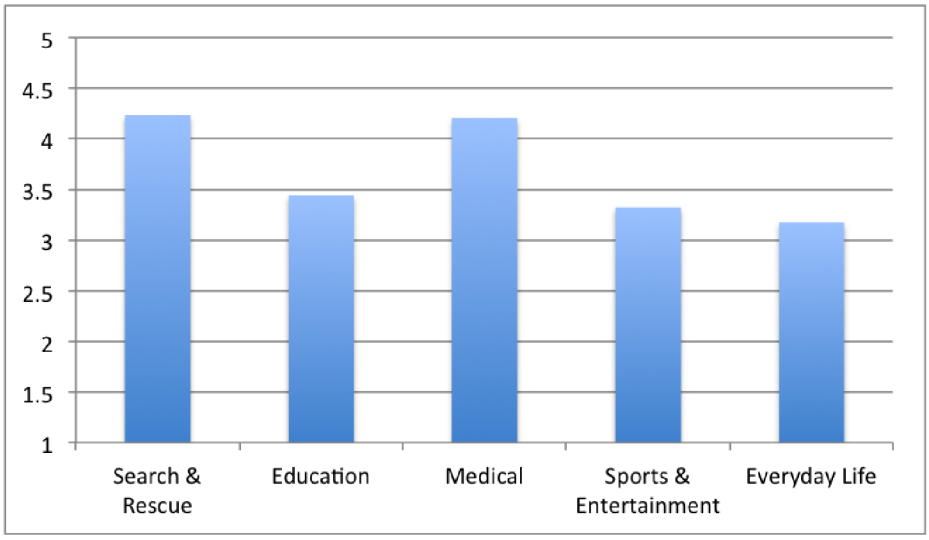


Figure 44: Post-survey results; Ethical Question: I can accept robot deception in [Context in x-axis] if it is strictly used only to benefit humans; Scales in y-axis: 1-strongly disagree, 5-strongly agree (recall from Figure 43); Average ratings to agree to other-oriented robot deception in specific context increase.

specifically provided, the subjects' acceptance rates increased. As shown in Figure 44, in search and rescue or medical situations, the average ratings of agreeing to use other-oriented robot deception increased to 4.23 ($\sigma^2 = 1.12$) and 4.20 ($\sigma^2 = 0.97$). From the results, I could initially formalize the ethical implications of robot deception as "the strong motives of deception in each context should be discussed and validated when other-oriented robot deception is used in an HRI context."

However, the results have limitations since the study participants were limited to older adult populations. In addition, there were only 39 participants in the study; this small number means their opinions cannot be generalized. To overcome these limitations, a web-based survey was conducted, and the opinions and impressions of robot deception were collected from the general public. The survey questions are identical to the post-survey questions in the previous HRI studies.

Participants were recruited using an email message. The email was sent to mailing lists that are found from campus, neighborhood, and online communities. The email message contains the link for the online survey, so if anyone wants to participate in

the survey, they can directly move to the survey page by clicking the link. A web-based consent form was gathered from participants. Once they open the survey link, a web-based consent form is shown in the first page. Participants can proceed to the survey questions only after accepting this form. Participants' agreement about the consent form is collected by pushing the "next" button on the bottom of the web-based consent form. Any participants who can legally agree using the online consent form are eligible for this study. Therefore, only those persons who are 18 years of age and older are asked to participate in the study.

The captured screens of this web survey are attached in Appendix C.2. The survey questions are identical to the ethical surveys from the HRI studies and the questions aim to access subjects' opinions on the use of robot deception in HRI. The survey made several ethical statements and the response was a rating on a five-point Likert scale (the ratings ranged from 1-strongly disagree to 5-strongly agree). Questions asked broadly whether they would accept a robot's other-oriented deception.

Overall, there were 174 participants in the survey, but 11 results were excluded due to incomplete responses. Thus, a total of 163 subjects' responses were analyzed. Table 30 illustrates the demographic information.

Table 31 illustrates the overall results of this survey study. All 163 subjects' responses were averaged for each question, and the results are represented in this table. According to the results, the participants were **not** very nervous when facing or interacting with a robot (average rating of Q1: 2.03 ($\sigma^2 = 1.06$), average rating of Q2: 2.10 ($\sigma^2 = 1.09$)). Similar to the results of HRI studies, subjects generally disagreed with the statement of robot deception (average ratings of Q5: 2.92 ($\sigma^2 = 1.06$), Q7: 2.74 ($\sigma^2 = 1.18$), and Q8: 2.74 ($\sigma^2 = 1.08$)). In other words, these average ratings were less than 3 points, which demonstrates that the participants could not determine the ethical acceptability of robot deception with these broad and high-level statements. However, as shown in the average ratings of Q9, when the situations

Table 30: Demographic Information from a web-based Survey Study

Total: 163 (Female: 70, Male: 93) **Average Age: 35.98** (min: 18, max: 74, σ^2: 12.15)		
Technical level		
Yes	68	41.71%
Somewhat	60	36.80%
No	35	21.47%
Prior experience with robots		
Never	51	31.28%
Very limited interaction	65	39.87%
Interaction experience with military robots	7	4.29%
Interaction experience with industrial robots	9	5.52%
Interaction experience with entertainment or educational robots	27	16.56%
Interaction experience with humanoid robots	3	1.84%
Others	1	0.61%
Age group		
Under 30	55	33.74%
31-40	53	32.51%
41-50	29	17.79%
Over 50	26	15.95%

Table 31: Survey Results Overview (1-strongly disagree, 5-strongly agree)

Questions about Robot Intelligence	
Question	Average (σ^2)
Q1: I would feel nervous just standing in front of a robot	2.03 (1.06)
Q2: I would feel nervous interacting with a robot	2.10 (1.09)
Q3: If something bad happens, I might depend on robots too much	2.46 (1.14)
Q4: If a robot's intelligence became equal to a human's in the future, I would accept it	3.08 (1.19)
Questions about Robot Deception	
Question	Average (σ^2)
Q5: A robot can hide/misrepresent information if it can help humans	2.92 (1.06)
Q6: The robot should always be honest in any circumstance	3.88 (0.96)
Q7: If humans can get benefits from robot's deceptive behavior, it can be accepted	2.74 (1.18)
Q8: Robot can intentionally/unintentionally deceive humans if it's in an appropriate situation	2.74 (1.08)
Q9: I can accept robot deception if it is strictly used only to benefit humans in the following context.	
Question	Average (σ^2)
Search and Rescue	3.41 (1.21)
Education	3.41 (1.21)
Medical (Rehabilitation)	3.44 (1.13)
Sports or Entertainments	3.2 (1.26)
Everyday life	2.85 (1.23)

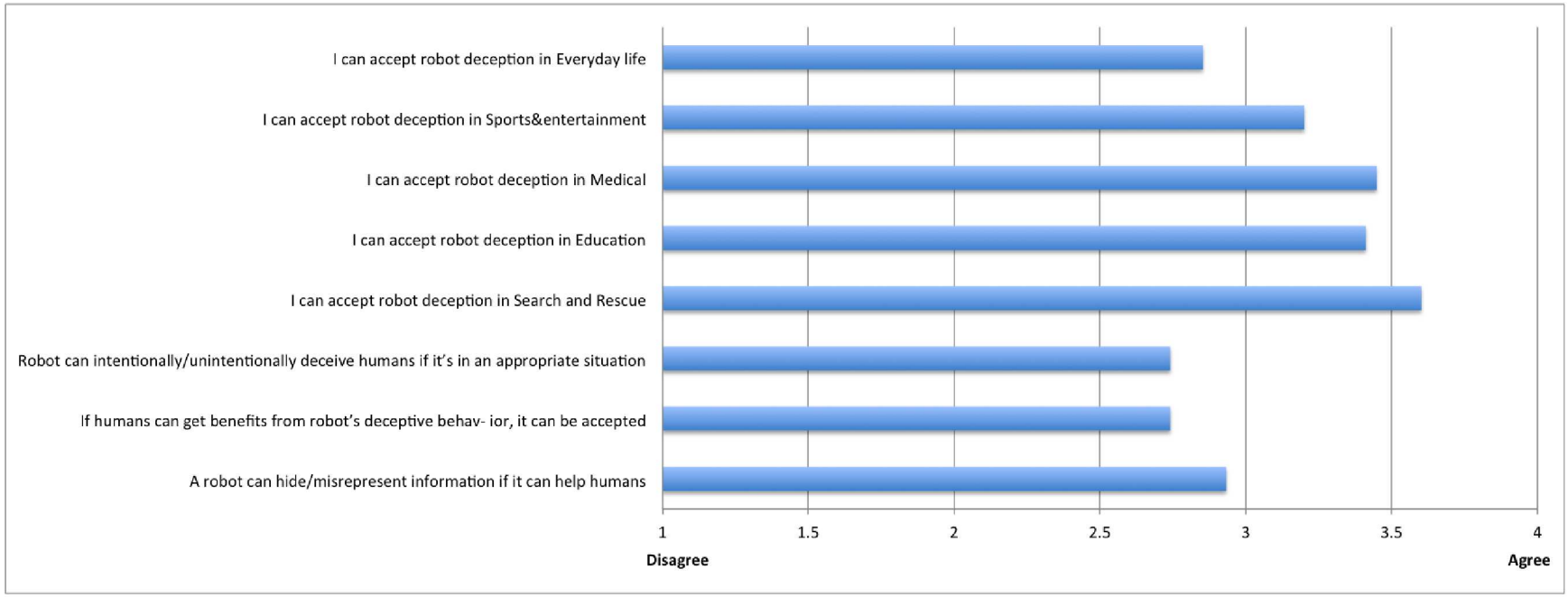


Figure 45: Survey Results Overview: I can accept robot deception in [Context in x-axis] if it is strictly used only to benefit humans; Scales in y-axis: 1-strongly disagree, 5-strongly agree

(contexts) were specified, the acceptance rates of robot deception slightly increased (Figure 45). Here, the survey statement was: "I can accept robot deception if it is strictly used only to benefit humans in the following context." From the literature reviews of human's other-oriented deception, five different contexts were selected as shown below.

- Search and Rescue: Robot rescuer hides current situation to human victims to calm down their panic level.
- Education: Robot assistant deceptively reacts to the students performance to motivate them and increase their learning efficiency.
- Medical (Rehabilitation): Placebo effects; Robot caregiver lies to patients if it can encourage them to accomplish more during the rehabilitation task.
- Sports or Entertainments: Robot soccer player fakes out opponents, thereby redirecting the opponents' actions.
- Everyday life: White lies.

These results were similar to those of the HRI study (chapter 5). When a specific situation (context) was provided, the subjects' acceptance rates slightly increased. In other words, these results support the ethical implication that I established in the previous chapter, which is that "People can accept the use of other-oriented robot deception when an appropriate and specific context is clearly determined." In sum, the strong motives of deception in each context should be discussed and validated when other-oriented robot deception is used in the HRI context.

The variations in demographic information were used to further analyze the results. First, it was observed whether or not people's acceptance of robot deception is different between age groups. Table 32 shows the survey's results of responses by age groups. The responses were divided into four age groups: under 30, 31 to 40,

Table 32: Survey Results by Age Groups (1-strongly disagree, 5-strongly agree)

Questions about Robot Intelligence				
Age Group	Q1	Q2	Q3	Q4
under 30	1.98 (1.02)	2.05 (1.07)	2.36 (1.02)	3.36 (1.19)
31 - 40	2.17 (1.14)	2.23 (1.16)	2.63 (1.25)	2.84 (1.24)
41 - 50	2.10 (1.11)	2.20 (1.17)	2.41 (1.05)	3.0 (1.15)
over 50	1.68 (0.80)	1.72 (0.79)	2.36 (1.31)	3.04 (1.09)
Questions about Robot Deception				
Age Group	Q5	Q6	Q7	Q8
under 30	3.23 (0.98)	3.80 (1.07)	2.98 (1.07)	3.0 (1.12)
31 - 40	2.82 (1.11)	3.90 (0.77)	2.80 (1.15)	2.66 (0.97)
41 - 50	2.75 (1.09)	3.89 (0.97)	2.32 (1.30)	2.51 (1.08)
over 50	2.60 (0.95)	4.04 (1.09)	2.52 (1.15)	2.56 (1.19)

Q9: I can accept robot deception if it is strictly used only to benefit humans in the following context.					
Age Group	SAR	Education	Medical	Sports and Entertainment	Everyday life
under 30	3.69 (1.16)	3.69 (1.19)	3.41 (1.19)	3.39 (1.30)	3.05 (1.31)
31 - 40	3.51 (1.03)	3.44 (1.22)	3.67 (1.09)	3.19 (1.22)	2.80 (1.22)
41 - 50	3.62 (1.11)	3.17 (1.07)	3.37 (1.11)	2.96 (1.26)	2.62 (1.20)
over 50	3.52 (1.19)	2.96 (1.27)	3.16 (1.10)	3.12 (1.26)	2.8 (1.15)

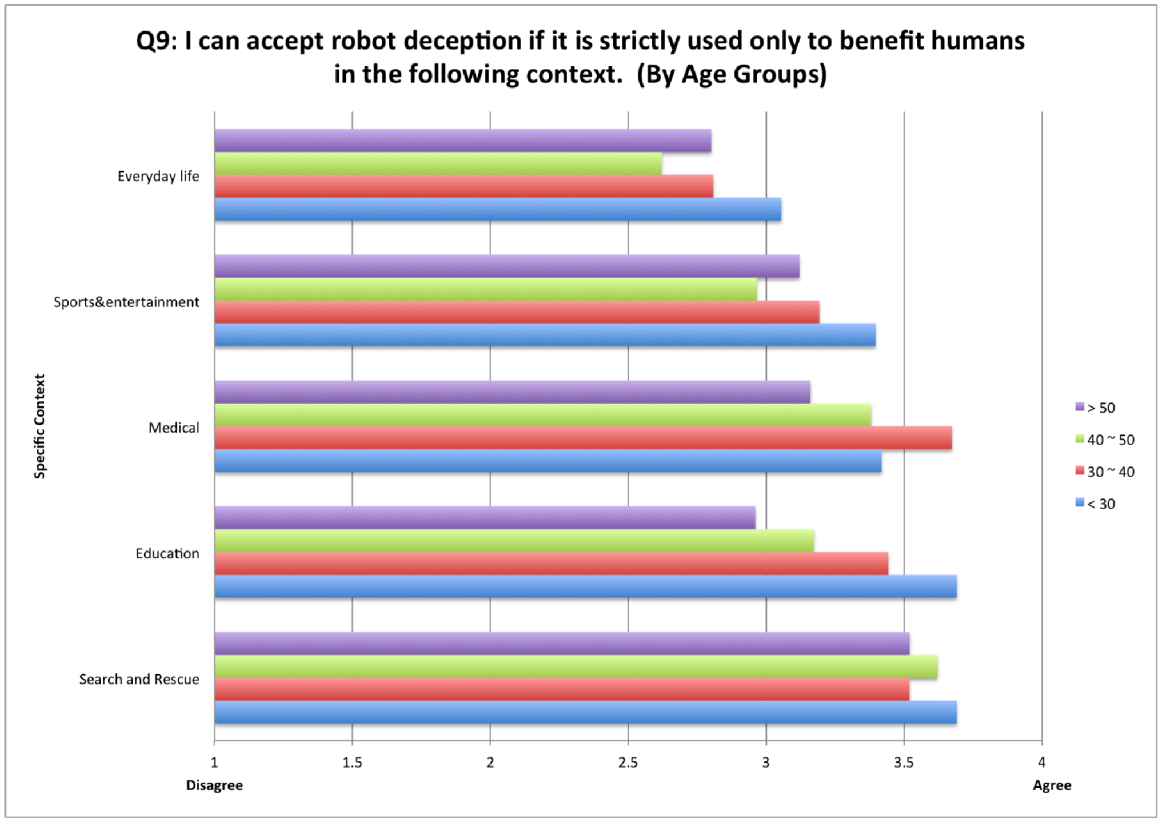


Figure 46: Survey Results by Age Groups: I can accept robot deception in [Context in x-axis] if it is strictly used only to benefit humans; Scales in y-axis: 1-strongly disagree, 5-strongly agree

Table 33: Survey Results by Technical Levels (1-strongly disagree, 5-strongly agree)

Questions about Robot Intelligence				
Technical?	Q1	Q2	Q3	Q4
Yes	2.08 (1.07)	2.14 (1.05)	2.48 (1.11)	3.26 (1.19)
Somewhat	1.91 (1.01)	1.91 (1.06)	2.31 (1.12)	3.11 (1.14)
No	2.14 (1.14)	2.34 (1.21)	2.65 (1.23)	2.68 (1.20)
Questions about Robot Deception				
Technical?	Q5	Q6	Q7	Q8
Yes	3.01 (1.08)	3.76 (1.03)	2.94 (1.26)	2.86 (1.15)
Somewhat	2.98 (1.01)	3.91 (0.82)	2.81 (1.08)	2.81 (0.98)
No	2.68 (1.07)	4.08 (1.03)	2.22 (1.08)	2.37 (1.08)

Q9: I can accept robot deception if it is strictly used only to benefit humans in the following context.					
Technical?	SAR	Education	Medical	Sports and Entertainment	Everyday life
Yes	3.57 (1.06)	3.42 (1.18)	3.57 (1.06)	3.43 (1.18)	2.91 (1.27)
Somewhat	3.7 (1.10)	3.46 (1.11)	3.6 (1.09)	3.06 (1.26)	2.95 (1.15)
No	3.48 (1.19)	3.28 (1.42)	2.94 (1.23)	3.0 (1.37)	2.57 (1.26)

41 to 50, and over 50. As shown in the table, there was no significant difference among age groups across all of the questions. One interesting finding was that the older groups ("41-50" and "over 50" groups) showed slightly low acceptance of robot deception in all of the contexts (Q9 in Table 32). As shown in Figure 46, among the five contexts, the difference between the younger and older groups was largest in the educational context (average rating from the "under 30" and "31-40" groups: 3.56 vs. average rating from the "41-50" and "over 50" groups: 3.06) and smallest in the search and rescue context (average rating from the "under 30" and "31-40" groups: 3.60 vs. average rating from the "41-50" and "over 50" groups: 3.57).

We can hypothesize that subjects' technical level and prior experiences with robots could affect their opinion on robot deception. To investigate this idea, the data was analyzed by technical levels and prior robot experiences. Table 33 and 34 contain the

Table 34: Survey Results by Prior Robot Experience (1-strongly disagree, 5-strongly agree)

Questions about Robot Intelligence				
Robot Experience?	Q1	Q2	Q3	Q4
Yes	1.81 (1.06)	1.89 (1.04)	2.44 (1.11)	3.46 (1.26)
No	2.12 (1.05)	2.18 (1.11)	2.46 (1.16)	2.93 (1.12)
Questions about Robot Deception				
Robot Experience?	Q5	Q6	Q7	Q8
Yes	3.10 (1.20)	3.68 (1.02)	3.19 (1.10)	3.02 (1.07)
No	2.86 (0.99)	3.97 (0.93)	2.56 (1.17)	2.62 (1.07)

Q9: I can accept robot deception if it is strictly used only to benefit humans in the following context.

Robot Experience?	SAR	Education	Medical	Sports and Entertainment	Everyday life
Yes	3.55 (1.19)	3.78 (1.12)	3.72 (1.13)	3.53 (1.27)	3.34 (1.27
No	3.62 (1.07)	3.25 (1.21)	3.33 (1.12)	3.06 (1.24)	2.65 (1.16)

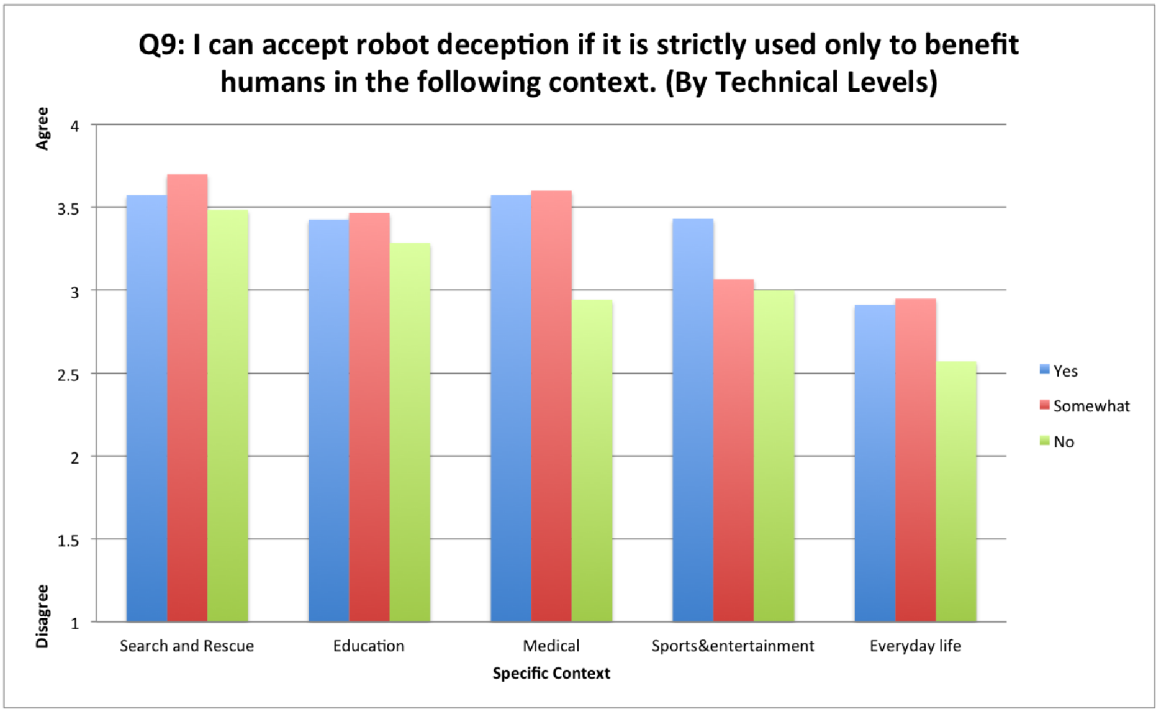


Figure 47: Survey Results by Technical Level: I can accept robot deception in [Context in x-axis] if it is strictly used only to benefit humans; Scales in y-axis: 1-strongly disagree, 5-strongly agree

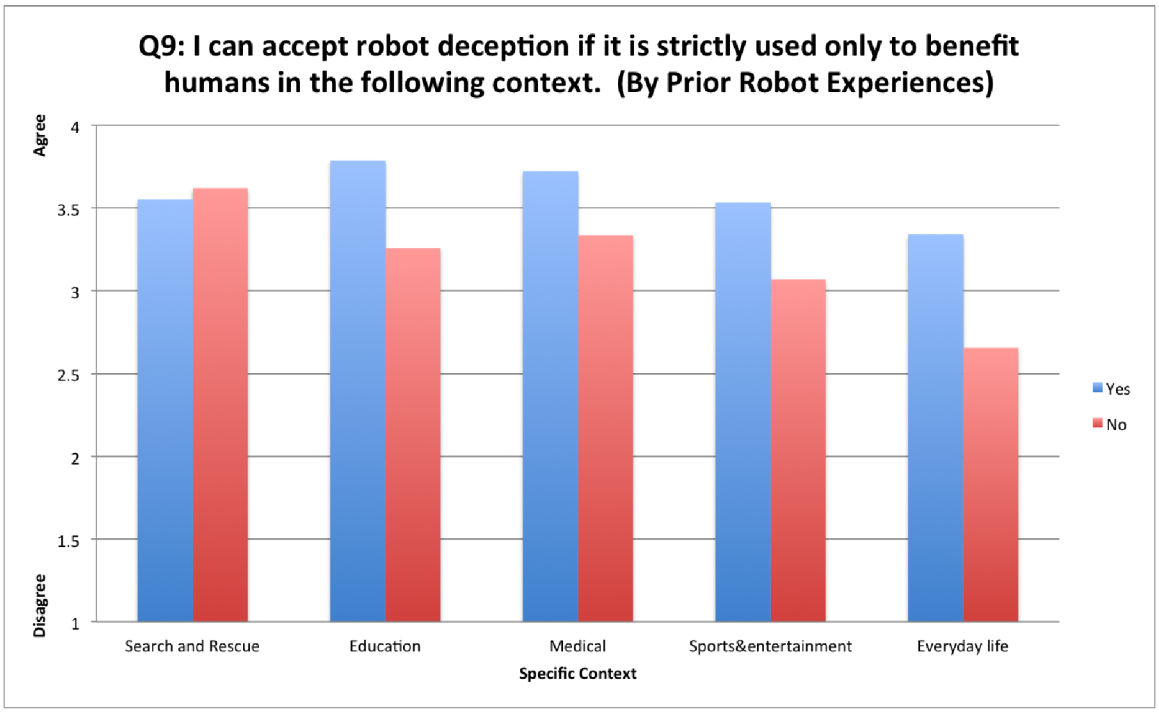


Figure 48: Survey Results by Prior Robot Experiences: I can accept robot deception in [Context in x-axis] if it is strictly used only to benefit humans; Scales in y-axis: 1-strongly disagree, 5-strongly agree

results of this analysis. Again, there are no significant differences in the acceptance of robot deception between different technical levels or prior robot experiences. However, as shown in the table, the subjects who had prior interaction experiences with robots had acceptance rates of robot deception slightly higher than other groups (Figure 48). The differences between two groups were larger in the everyday life, educational, and sport/entertainments contexts, whereas the difference was significantly smaller in the search and rescue context. From these results, I can also assume that when the context is a more life-threatening situation, the acceptance rates are similarly higher in all of the groups.

6.3 Summary

Robot deception is an ethically sensitive topic, and extensive ethical discussion is required for this research. In fact, deception is an ethically arguable behavior even

when performed by humans. However, it is also true that deceptive capabilities include potential benefits, and by carefully applying this capability to the robotic system, we can achieve more socially intelligent robots in human-robot interaction contexts.

To discuss ethical argumentation related to robot deception, especially other-oriented robot deception, the literature was reviewed and and survey research were conducted. First, post-survey responses were gathered from HRI study participants, as described in Chapter 5. However, these results had limitations since it only included 39 participants from the target population (older adults group). To overcome these limitations, a web-based survey research was conducted and 163 responses were gathered from the general population.

The online survey results revealed that, in general, people do not accept robot deception, even though it is described as other-oriented robot deception. However, when the specific contexts or practical uses are clearly described, these acceptance rates can increase. From the analysis of online survey results, the following findings were illustrated in this chapter.

- An average ratings of agreeing to general concept of other-oriented robot deception is less than points 3 (undecided).
- An average ratings of agreeing to use other-oriented robot deception increased when an appropriate context is provided with justification.
- Related to demographic information, two features seem to affect human's acceptance to other-oriented robot deception, which are age and prior robot experience.
  - The older people showed slightly low acceptance of robot deception.
  - People who had prior interaction experiences with robots had acceptance rates of robot deception slightly higher than other groups.

- When the context is a more life-threatening situation and the use of other-oriented deception is clearly verified, the acceptance rates are higher than other situations.

I argue that this ethical implication highly correlates to the motives of robot deception. As explained before (Chapter 4), I formulated the robot deception mechanism into three dimensions; motive, opportunity, and method. The ethical implications determined from this survey argue that the context should be clearly validated. In terms of my theory, validating (from domain experts) whether or not the context includes strong **motives** for robot deception is essential for achieving other-oriented robot deception. In sum, I argue that when other-oriented robot deception is used in a practical situation, experts' validation for the motives should be prioritized.

CHAPTER VII

CONCLUSION

Deception is an ethically arguable behavior, but at the same time it is essential social behavior for humans. We can observe human deceptive behaviors in a variety of contexts including sports, culture, education, war, and everyday life. Deception is not only limited to human beings. It is commonly used for the purpose of survival in animals and even in plants. From these findings, it is obvious that deception is a general and essential behavior for any species, which raises an interesting research question: can deception be an essential characteristic for robots, especially social robots?

Based on this curiosity, I aimed to develop a robot's deception capabilities, especially in human-robot interaction situations. In addition, I strongly argued that a social robot's deceptive behaviors should only be used when it can produce benefits for the deceived humans. As a result, my primary research question for this dissertation was formalized as follows: *Can a robot use deception in appropriate HRI domains in order to benefit the deceived human partner?*

To prove my primary research hypothesis, as well as to achieve this benevolent robot deception, I broke my research down into five subsidiary questions. Throughout this dissertation, I answered the subsidiary questions as follows.

1. *What kinds of deception can be beneficial for those being deceived?*

   Objective: Humans and animals can use different kinds of deceptive behaviors. Similar to humans and animals, robots can also perform deceptive behaviors for specific purposes. Among different contexts, a number of potential situations in

which, deceptive robot behaviors can be beneficial, should be investigated. The aim of this research is to determine the appropriate use of robots' deceptive behaviors to benefit the deceived. Therefore, it is essential to understand in which particular contexts robot deception may benefit deceived people.

Result and Contribution: Literature related to deception in a variety of fields was reviewed in chapter 2. Deception in biology (section 2.1) and psychology (section 2.2) was highlighted, and previous work in robot deception was extensively reviewed (section 2.3). Particularly, the definition of other-oriented deception in psychology was introduced and also situational conditions for other-oriented deception in humans were represented. To validate and determine the potential deceptive capabilities for a robot to benefit the deceived humans, robot deception needs to be categorized. However, there is a lack of studies on basic knowledge and fundamental theory in robot deception. To overcome this limitation, I reviewed previous research on deception and developed a novel taxonomy for classification of robot deception as a preliminary work (chapter 3). From this taxonomy, I defined the terminology of "other-oriented robot deception." Several clues on why deceptive behaviors are essential in a robotic system were found. In addition, from the categorization of deception, potential contexts in which robot deception can benefit the human being deceived were discussed.

2. How can deceptive behaviors be applied to a robotic system?

Objective: Deceptive behaviors have to be applied appropriately to robotic systems. Since robots differ from animals and humans in their embodiment and motion/perception capabilities, it will be necessary to determine the most applicable methodologies for robot systems under these limitations and conditions.

Result and Contribution: When applying behaviors or actions to a robotic system, a robot's capabilities and platform characteristics should be considered. In my model, these factors were considered when developing a deceptive action generation

model (section 4.2). Since this is one part of my computational model, it will be more specifically described with the answer to question 3 below.

3. What formal theoretical/mathematical expressions are appropriate for generating robot deception?

Objective: Algorithms should be developed to apply deceptive behaviors to the robot system. Formal theoretical expressions and suitable computational models require development. In particular, I am concerned about the deceptive capabilities of robots in HRI contexts. Therefore, the development of formal deceptive expressions for a robot while interacting with people is necessary for this research.

Result and Contribution: After arguing the importance of other-oriented robot deception, the research aimed to achieve a model for a robot to perform other-oriented deception. This was solved by answering subsidiary question 2 and 3. To achieve other-oriented robot deception, a computational model for a robot's other-oriented deception needed to first be developed (Question 3) and implemented into an appropriated robot platform (Question 2). In chapter 4, a novel computational model for a robot's other-oriented deception was presented.

The model is inspired by criminological definition of deception. According to criminological findings, deception is analyzed by three criteria, which are motive, method, and opportunity. Similar to this approach, in my model a robot first has to determine whether the current situation includes any motives to perform the deceptive behaviors. If so, then a robot should generate the methods to perform deception. Finally, by selecting among different true/deceptive behaviors, it should be possible to determine which one is the most appropriate in a certain situation, thus providing opportunity. According to this approach, the method model was developed; a deceptive action generation mechanism inspired by Bell and Whaley's deception categorization (section 4.2). Since the robot's behavior or action depends on the robot platform's features,

the action generation model also included integration and filtering steps (Question 2). Then, as the motive and opportunity model, deceptive action selection mechanism is generated via case-based reasoning model (section 4.3). Finally, by integrating those models together, the computational model for a robot's other-oriented deception can be achieved (Question 3). To show how the model works, this computation model was also reviewed with a specific example at the end of chapter (section 4.3.2).

4. What are the most effective evaluation methods and metrics to test the research hypothesis?

Objective: After the computational models for generating robot deception are determined and applied to robots, the algorithm must be tested to evaluate if it is truly working. Furthermore, in this dissertation, I aim to address my research hypothesis, which is that robots' deceptive behaviors can benefit deceived people in certain HRI contexts. The hypothesis must be tested to determine whether it is correct according to the specific developed deceptive behaviors for the robot. To answer these questions, we have to conduct well-designed HRI studies with human subjects as evaluation methods.

Result and Contribution: After successful implementation, the research hypothesis needed to be tested and proved via appropriate HRI studies. Chapter 5 discussed several HRI studies and their results. As a study design process, it was essential to validate an appropriate HRI context by observing human-human cases. For this purpose, I reviewed multiple situations where humans frequently used other-oriented deception, and finally selected elderly persons' rehabilitation as an HRI study context.

The design of the HRI study was inspired by elderly's rehabilitation. In particular, elderly's motor-cognition dual task was selected as the task. While the participants performed the assigned motor-cognition dual tasks, the robot assistant was placed next to them and provided feedback to their performance. Basically, the robot

provided positive feedback when participants correctly solve the tasks and negative feedback when the tasks were solved incorrectly. However, the robot was also capable of performing other-oriented deceptive feedback, and motivating and encouraging the participants. Occasionally, the robot also generated deceptive positive feedback even when the participants incorrectly answered the questions.

By conducting this HRI study, it was aimed to observe whether a robot's other-oriented deceptive feedback can help human subjects increase their performance in rehabilitation tasks. 34 participants who are over 55 years old were recruited and a two by two mixed-subject study was conducted. From the results, the average frustration level rating was significantly reduced in the deception group (avg. = 6.47) compared to the true group (avg. = 9.58). According to the self-reported measures on impressions of a robot or monitor feedback, the average ratings of task enjoyment were also significantly higher in the robot feedback group (avg. = 4.26) compared to the monitor feedback group (avg. = 3.22). Therefore, it can be concluded that a robot's other-oriented deception can potentially reduce the participant's frustration and increase the participant's enjoyment during the rehabilitation tasks.

To test the deceptive action selection model, an extended HRI study was also conducted. The HRI study procedures were the same, except the robot's action selection was made by CBR-based deceptive action selection model. From the results, it was observed that the model is performed appropriately with real human subjects. Finally, chapter 5 demonstrated the successful use of other-oriented robot deception and the potential benefits for the deceived humans in an appropriate context, as posed in subsidiary question 4.

5. How should the ethical issues of robot deception be handled in HRI?

Objective: Even though robot deception can provide several advantages to humans, it is arguable whether deceiving humans is morally acceptable in HRI. I will

also consider this ethical issue thoughtfully in this research.

Result and Contribution: Using deceptive behaviors obviously leads to ethical arguments, even in human cases. Therefore, a discussion of ethical issues in robot deception is an essential part of this research. For this consideration, I also reviewed moral theories and ethical approaches to deception in section 2.4. In addition, to further discuss ethical argumentation related to other-oriented robot deception I conducted a survey research as represented in chapters 5 and 6. First, post-survey responses were gathered from the HRI study participants. However, this result had limitations since it only included 39 participants from the target population (older adults group). To overcome this limitation, I extended this post-survey and conducted a web-based survey research. As a result, I gathered 163 responses from the general population, which can lead to generalized ethical implication for robot deception.

The survey results revealed that people do not accept robot deception (even though it is described as other-oriented robot deception) in general. However, when the specific contexts or practical usages are clearly described, people's acceptance rates can increase. I can argue that this ethical implication correlates to the motives of robot deception. As explained before, I formalized robot deception mechanism in three dimensions, which are motive, opportunities, and methods. The ethical implications found from this survey argues that the context should be clearly validated.

By answering the five subsidiary questions, I addressed my primary research hypothesis. The benefits of robot deception in an appropriate HRI situation was demonstrated. In sum, from the results of the primary and subsidiary research questions, the following scientific contributions were accomplished in this dissertation:

- A novel taxonomy of robot deception is defined based on significant literature reviews on deception in a variety of fields, such as psychology, biology, military, economics, and so on.

- A general computational model for a robot's deceptive behaviors is developed based on criminological law.
- Appropriate HRI contexts in which a robot's other-oriented deception can generate benefits are explored and assessed.
- A methodology for evaluating a robot's other-oriented deception in appropriate HRI contexts is designed, and studies are conducted with human subjects.
- The ethical implications of other-oriented robot deception are explored and thoughtfully discussed.

There remain many arguments and limitations to using robot deception in HRI contexts broadly and practically. However, as shown in this dissertation, other-oriented robot deception can potentially produce benefits to humans, and therefore, the use of other-oriented deception should be considered for social robots in HRI. Finally, I emphasize that this dissertation defined and achieved other-oriented robot deception, so it can potentially contribute to developing a benevolent deceptive social robot for human-robot interaction.

APPENDIX A

EXPERIMENTAL SETTINGS OF SQUIRREL ROBOT DECEPTION

A.1 Essential States and Triggers in MissionLab

To implement and evaluate the squirrel robot deception model, MissionLab is used.

General States and Triggers

State/Trigger	Descriptions
GoTo (parameter: location x, y)	The robot is move to the parameterized location (x,y)
Wander	The robot is wandering around the specified map.

State and Triggers for Squirrel Robot

State/Trigger	Descriptions
DetectColorBlob (parameter: color)	This trigger is activated when the vision server detects the parameterized color (i.e., isBlobDetected() in the vision server is activated).
StayUntil	The robot stop until n seconds, where n is determined by the number of cached items.
Prob_con	The robot determines the distribution of transition probabilities of the current locations and it is probabilistically likely to select one of the location among current places.
Prob_loc	The robot is probabilistically likely to select one of the location among current places.

State and Triggers for Competitor Robot

State/Trigger	Descriptions
Prey_found (parameter: color)	This trigger is activated when the squirrel robot is detected from the vision server. This trigger is the modification from DetectColorBlob trigger.
Notified_ObjectLocation	When the competitor robot receives the location information of the detected prey from the vision server, it processes to the next step.
Exceed_Threshold (parameter: time)	If the maintaining time of the notified object is over the parameterized time (i.e., isOverThreshold() in the vision server is activated), this trigger is activated.

A.2 Probabilistic Transition

Transitions based on the existence of probabilities to simulate environmental properties such as number of cached items. First, P_{ij} is the transition probability that indicates the location j is selected as the next patrol location when the current location is in location i. In addition, n is the total number of locations and $\#item_x$ indicates the number of food items in location x.

$$p_{ij} = \frac{\#items_j}{\sum_{l \leq k \leq n, k \neq j} \#items_k} \tag{15}$$

Based on the determined probabilities, a weighted roulette wheel algorithm is applied to decide the next transition. Using the transition probabilities, the proportion of the wheel is assigned to each of the possible selections. Then, if the randomly generated number is fitted to one of the proportion, it decides as the next transition:

$$R = \text{random number between 0 and 1}$$
$$NextLocation_i = \begin{cases} Location_1, & R \leq P_{i1} \\ Location_2, & R \leq P_{i1} + P_{i2} \\ \vdots & \end{cases} \tag{16}$$

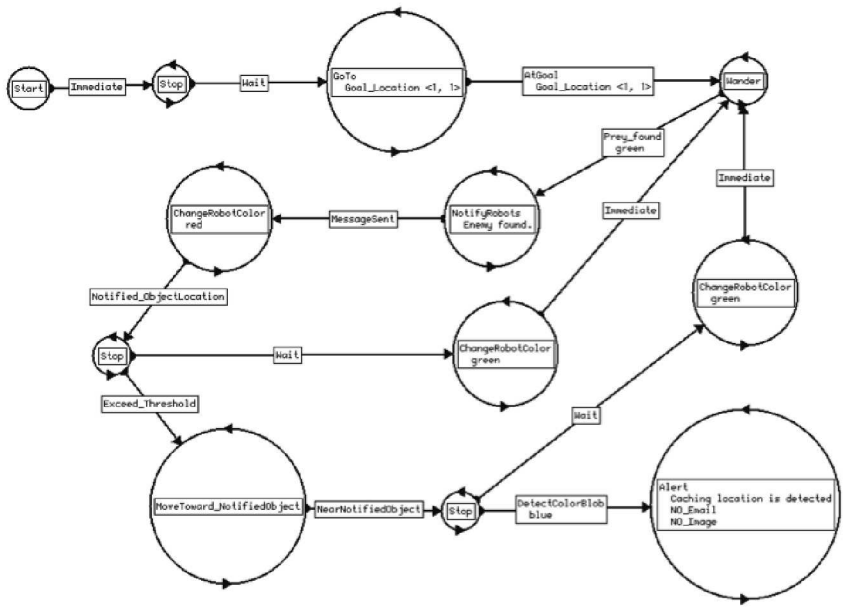


Figure 49: Finite State Acceptors for competitor robot's hunting strategy in MissionLab

A.3 Implementation of Competitor Robot

A competitor robot has a simple strategy in the experiment. Here, the competitor robot wanders around the map to try and find the squirrel robot. When it detects the quarry, it determines whether it is at a potential caching location or not, by observing how long the squirrel robot stays in place. Since the squirrel robot takes time to patrol the caching place proportional to the number of food items, the competitor robot obtains evidence of the caching area based on the robot?s time onsite. Therefore, if the duration is over a threshold, set empirically, the competitor robot recognizes the place as a caching area. The competitor robot then goes to the detected location and pilfers. In our robot implementation, in the pilferage step, the competitor robot confirms whether the detected location truly contains the items by discriminating food items based on the colors. If it confirms the caching location, it sends the alert message to the system that it has found the cached item. If it determines the location doesn't include the caching item, it returns to "wander" state and repeats the detecting process again. Figure 49 shows the FSA of the competitor robot in

the real robot experiment. As shown in this figure, the robot used GoTo, Wander, MoveTowardObject, Detect, and Alert behaviors for the competitor robot?s strategy.

A.4 Parameters used in the real robot experiments

Squirrel Robot's parameter setting in MissionLab

Move_to_location gain	0.9
Wander_gain	0.0
Avoid_obstacle_gain	0.9
Avoid_obstacle_sphere	0.7
Avoid_obstacle_safety_margin	0.2
Max_velocity	1.5
Base_velocity	1.5

Competitor Robot's parameter setting in MissionLab

Move_to_location gain	0.9
Wander_gain	0.5
Avoid_obstacle_gain	0.9
Avoid_obstacle_sphere	0.7
Avoid_obstacle_safety_margin	0.2
Max_velocity	1.0
Base_velocity	1.0

APPENDIX B

HRI STUDY SUPPLEMENTS

B.1 Study Flyer

Georgia Tech College of Computing
Defining the New Face of Computing

STUDY PARTICIPANTS WANTED!!!

The Georgia Tech Mobile Robot Lab is conducting a human-robot interaction experiment and YOU have a chance to interact with a humanoid robot Nao!

The study will take place in Tech Square (TSRB building) at Georgia Tech, and you will be asked to perform interesting motor-cognition tasks with a NAO robot while we're gathering your physiological data using E4 wristband (watch-like wearable device).

This study is designed for elderly, so you must be ***at least 50 years old*** in order to participate. You have to have English language proficiency and should be physically free to move, read, and listen. This study will take no more than 60 minutes of your time. You don't need to have any technical knowledge to participate in this study, and we invite all backgrounds to partake.

As a token of our appreciation, you will receive $5~15 cash based on your task achievement. Participate and earn the maximum prize!

If you are interested in participating, please contact Jaeeun Shim via email (**jaeeun.shim@gatech.edu**) or phone (**404-831-1660**).

B.2 Consent Form

GEORGIA INSTITUTE OF TECHNOLOGY
CONSENT TO BE A RESEARCH PARTICIPANT

Project Title: Human-Robot Interaction Study

Investigators: *Arkin, R.C., Ph.D. and Shim, J.*

Protocol and Consent Title: *The Benefits of a Robot feedback during a motor-cognition dual task consent form for adult participation*

You are being asked to be a volunteer in a research study.

Purpose:
The purpose of this study is to evaluate how a robot behaves when it is interacting with humans. The results will help improve the design of robots and promote robotics research. We expect to enroll 40-50 people in this study.

Exclusion/Inclusion Criteria:
Participants in this study must be 50 years old or older and be able to use English at a high school level. Also, participants should be physically able to move, read, and listen.

Procedures:
You will be invited to Georgia Tech's Mobile Robot Laboratory, where you will first complete a questionnaire. After that, you will be asked to perform a task in two different conditions.

The task is sorting weekly pills. While you're doing the pill-sorting tasks, some questions will randomly be asked and you will try to answer.

In one of the task's study conditions, a small humanlike robot will be placed next to you and provide feedback. The robot will respond whether or not you answer the questions correctly, by giving you a happy or sad gesture.

In the other study conditions, instead of the robot, a computer monitor provides feedback using a green or red color. Simply, it will show you a green color if you answer correctly and a red color if you answer incorrectly.

The total study time should not exceed 60 minutes.

Consent Form Approved by Georgia Tech IRB: March 14, 2016 - February 18, 2017

During the study, we will videotape your performance. With your permission, we will also gather your skin temperature, heart rate, and blood pressure using a wearable device (E4 wristband). At the end of the study, you will be asked to fill out another questionnaire. This questionnaire will ask you about your impressions and opinions of the robot.

You can stop at any time during the study for any reason. During the study, you may be led to believe some things that are not true. When the study is over, we will tell you everything. At that time you can decide whether to let us use your information. You have the right to then require that your information be destroyed and not be used in the study.

Risks or Discomforts:
This study should not put you at risk. The risks involved are no greater than those involved in playing a video or a role-playing game.

Benefits:
You are not likely to receive any significant benefits from joining this study. However, you will have an opportunity to interact with a humanlike robot and expand your knowledge about robots in general.

Compensation to You:
For your time and effort, you will be compensated with between $5 and $15 cash.

Confidentiality:
We will follow some procedures to protect your personal information that you share in this study: The data collected about you will be kept private to the extent allowed by law.

To protect your privacy, your records will be kept under a code number rather than by name. Your records, including videotapes, will be kept in locked files and only the study's staff will be allowed to look at them. The videotapes will be destroyed after we analyze the data. Your name and personal information will not appear when we present or publish the results of this study. At the end of this form, we will ask for your written permission for use of any videos or photographs in our demos and publications.

To make sure that this research is being carried out in the proper way, the Georgia Institute of Technology Institutional

Consent Form Approved by Georgia Tech IRB: March 14, 2016 - February 18, 2017

Review Board (IRB) may review study records. The Office of Human Research Protections may also look over study records during required reviews.

Costs to You:
There are no costs to you, other than your time, for being in this study.

In Case of Injury/Harm:
If you are injured as a result of being in this study, please contact Principal Investigator, Ronald C. Arkin, Ph.D., at telephone (404) 894- 8209. Neither the Principal Investigator nor Georgia Institute of Technology has made provision for payment of costs associated with any injury resulting from participation in this study.

Participant Rights:
- Your participation in this study is voluntary. You do not have to be in this study if you don't want to be.
- You have the right to change your mind and leave the study at any time without giving any reason and without penalty.
- You will be given a copy of this consent form if you request it.
- You do not waive any of your legal rights by signing this consent form.

Conflict of Interest:
None.

Questions about the Study:
If you have any questions about the study, you may contact Principal Investigator, Ronald C. Arkin, Ph.D., at telephone (404) 894- 8209 or at arkin@cc.gatech.edu.

Questions about Your Rights as a Research Participant:

If you have any questions about your rights as a research participant, you may contact

Ms. Melanie Clark, Georgia Institute of Technology
Office of Research Compliance, at (404) 894-6942.

Consent Form Approved by Georgia Tech IRB: March 14, 2016 - February 18, 2017

If you sign below, it means that you have read the information given in this consent form, and you would like to be a volunteer in this study.

_____
Participant Name (printed)

_____ _____
Participant Signature Date

_____ _____
Signature of Person Obtaining Consent Date

Video Release:
With your permission, Georgia Tech Mobile Robot Laboratory may use the video footage or photographs from the videos, which contain your appearance. The photographs or video recordings will only be used for research or educational purposes. This may include conferences displays, briefings, workshops, etc, but not any commercial uses.

If you sign below, it means that you accept the conditions of the releasing your video as stated above.

_____
Participant Name (printed)

_____ _____
Participant Signature Date

_____ _____
Signature of Person Obtaining Consent Date

Consent Form Approved by Georgia Tech IRB: March 14, 2016 - February 18, 2017

B.3 Pre-survey forms

Participant # _____

Demographics Questionnaire

1. What is your gender?
   ☐ Female ☐ Male

2. What is your age? _____

3. What is the highest level of education you've achieved?
   ☐ High School ☐ Bachelor's ☐ Master's ☐ Ph.D. ☐ Other _____

4. Do you describe yourself as technical (having extensive experience or interest in a technical field, such as engineering, computing, math, etc)?
   ☐ Yes ☐ Somewhat ☐ No

5. What is your level of computer experience?
   ☐ None: Never used a computer before
   ☐ Limited: Occasionally use a computer for tasks like e-mail, internet or word processing
   ☐ User Level: Regularly use a computer for tasks like e-mail, internet or word processing
   ☐ Advanced User: Have downloaded and installed at least one program from the Internet
   ☐ Programmer Level: Some programming language or network administration experience
   ☐ Advanced Programmer: Extensive training or experience in programming languages

6. Have you ever interacted with robots? Please check all that apply.
   ☐ Never ☐ Very limited interaction
   ☐ Interaction experience with military robots
   ☐ Interaction experience with industrial robots
   ☐ Interaction experience with entertainment/educational robots
   ☐ Interaction experience with humanoid robots
   ☐ Other – please specify _____

7. Have you participated in another experiment with this robot in this lab?
   ☐ Yes ☐ No

8. Please choose all current and past medical conditions.

□ No medical problem	□ Heart failure	□ Liver disease	□ Hearing Loss
	□ Stroke	□ Osteoporosis	□ Deafness
□ Diabetes	□ Lung disease	□ Osteoarthritis	□ Cataracts
□ Bleeding disorders	□ Asthma	□ Cancer – where?	□ Glaucoma
□ High blood pressure	□ Bronchitis	_____	□ Macular Degeneration
□ Heart attack	□ Emphysema	□ Seizures	□ Blindness
□ Stomach ulcers	□ Kidney Failure	□ Anxiety	□ Color Blindness
□ Blood clots in legs/lung	□ Kidney Stones	□ Depression	□ Other: _____
□ Abnormal heart rhythm		□ Seen a psychiatrist	

Participant # _____

Indicate how much you agree/disagree with the following statements. Provide a rating for each statement based on the scale shown below.

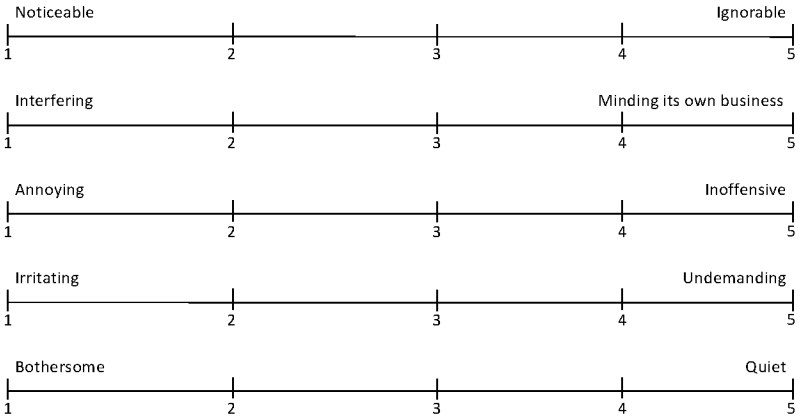
	Strongly disagree	Disagree	Undecided	Agree	Strongly agree
Something bad might happen if robots developed into living beings.	1	2	3	4	5
I would feel relaxed talking with robots.	1	2	3	4	5
I would feel uneasy if I was given a job where I had to use robots.	1	2	3	4	5
The word “robot" means nothing to me.	1	2	3	4	5
I would feel nervous operating a robot in front of other people.	1	2	3	4	5
I would hate the idea that robots or artificial intelligences were making judgments about things.	1	2	3	4	5
I would feel very nervous just standing in front of a robot.	1	2	3	4	5
I feel that if I depend on robots too much, something bad might happen.	1	2	3	4	5
I feel that in the future society will be dominated by robots.	1	2	3	4	5
The robot should always be honest in any circumstance.	1	2	3	4	5

B.4 Self-reported measures: Impressions to a robot feedback

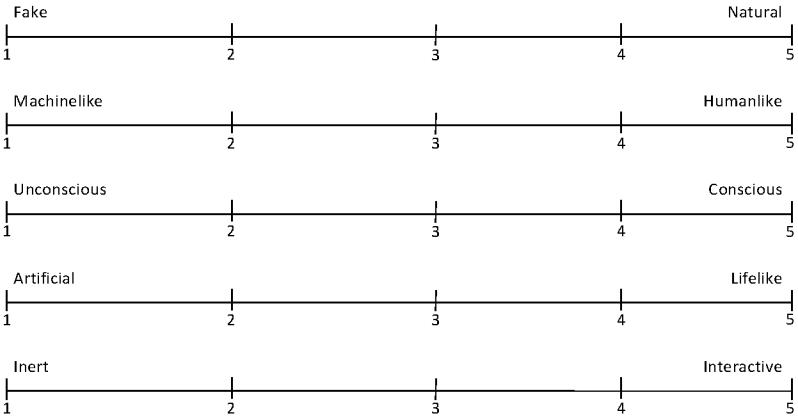
Participant #

Please reflect back on your interaction with the robot during the task you have just completed as you consider the questions below. **Please rate your impressions of the robot DURING THIS TASK by circling the most appropriate number on the scale:**

1. In your opinion, during this task, the robot was:

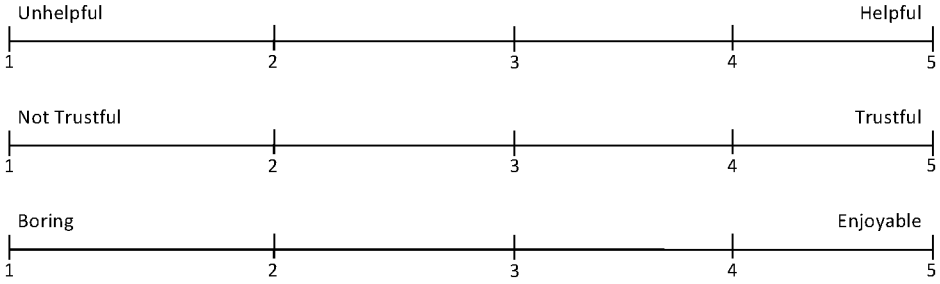


2. In your opinion, the robot APPEARED:



TURN OVER →

3. During the task, the feedback robot was:



4. In your own words, please briefly describe your interaction with the robot. Please mention if there were any changes in your thoughts or feelings throughout the interaction.

B.5 Self-reported measures: Impressions to a non-robotic visual feedback (monitor)

Participant #

Please reflect back on your interaction with the robot during the task you have just completed as you consider the questions below. **Please rate your impressions of the way to provide the feedback using the monitor screen. DURING THIS TASK by circling the most appropriate number on the scale:**

1. In your opinion, during this task, feedback from the monitor screen was:

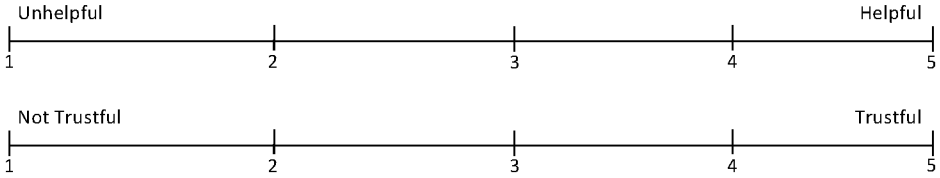
Noticeable	1	2	3	4	5	Ignorable
Interfering	1	2	3	4	5	Minding its own business
Annoying	1	2	3	4	5	Inoffensive
Irritating	1	2	3	4	5	Undemanding
Bothersome	1	2	3	4	5	Quiet

2. In your opinion, the monitor screen APPEARED:

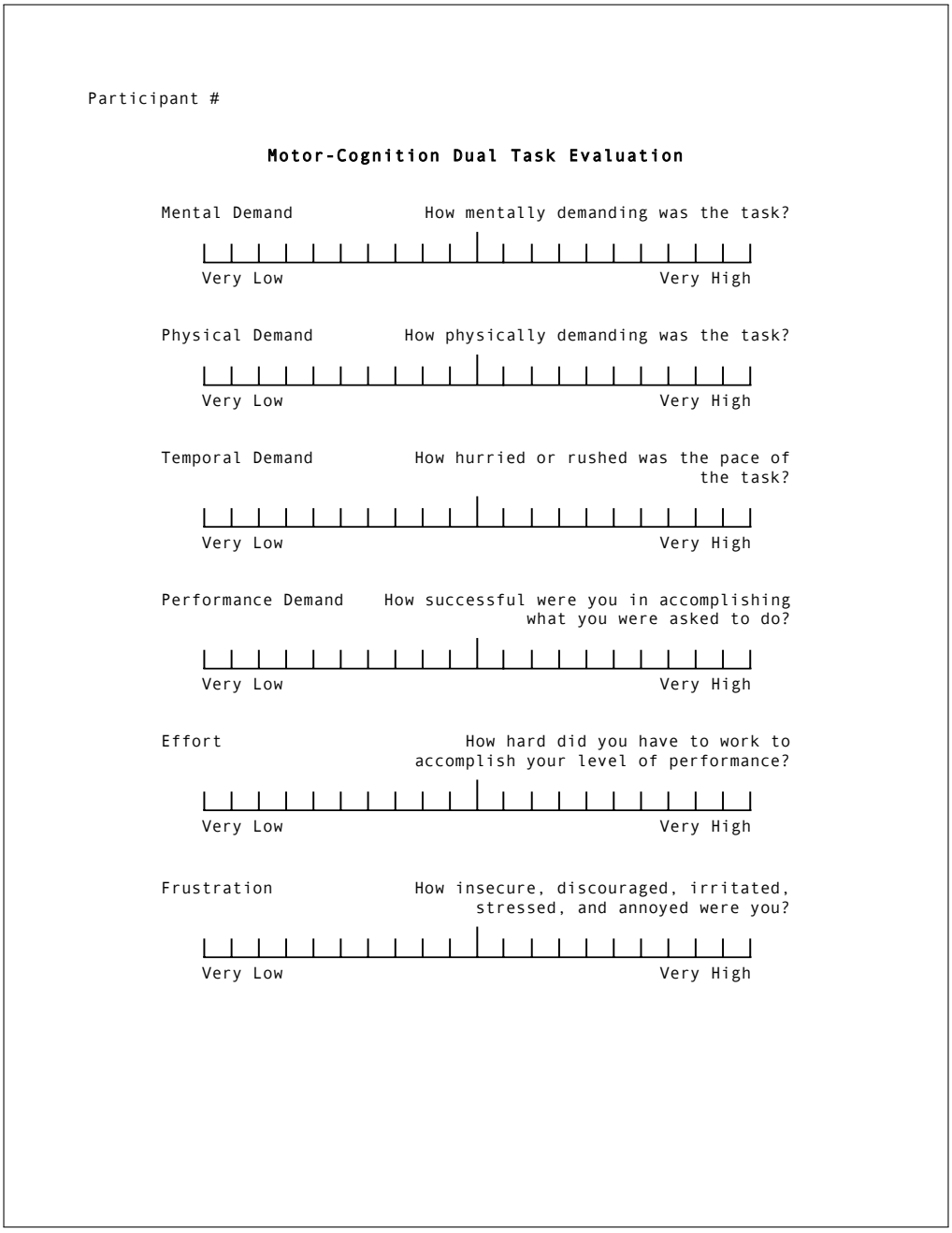
Fake	1	2	3	4	5	Natural
Machinelike	1	2	3	4	5	Humanlike
Unconscious	1	2	3	4	5	Conscious
Artificial	1	2	3	4	5	Lifelike
Inert	1	2	3	4	5	Interactive

TURN OVER →

3. During the task, feedback from the monitor screen was:



B.6 Self-reported measures: Workload impressions using NASA's TLX



B.7 HRI study debriefing forms

Study Debriefing Form

Study Debrief Information
This study is designed to observe a human's performance based on a robot's true or deceptive feedback. Therefore, the study consists of two conditions. In the true condition, a robot provides true feedback based on the subject's performance. In this condition, if the subject answers the auditory task correctly, the robot shows a positive gesture and if incorrectly, it shows a negative gesture. The other condition is deceptive feedback and here a robot provides feedback deceptively. More specifically, when the subject incorrectly answers more than two questions, the robot provides positive feedback even though the subject answers incorrectly, and so on.

You were in the **true condition**, in which the robot always provided you true feedback. Half of the other participants were placed in the deception condition, in which the robot or the monitor gave some deceptive feedback when they were performing the tasks.

Purpose of the Study
We aim to observe how a human's engagement and task performance can be improved or degraded in response to the feedback condition. In addition, in order to analyze the effects of robot's embodiment, we also run the study with a baseline condition, which is the non-robotic feedback device, the monitor.

Description of the Deception
There was one more hidden element that we didn't provide you at the beginning of this study. From the study, we need to observe how participants' motivations and engagements to the task will vary according to a robot's feedback. Therefore, the study was designed so that participants were motivated or unmotivated by the benefits. For this purpose, at the beginning of the study, we informed participants that their compensation would be differentiated by their performance according to the compensation guidelines. This was just an experimental method. You will receive the full amount of compensation, which is $15 regardless of your performance.

Purpose of the Deception
We aim to observe how participants' performance will vary responding to a robot's feedback. To make participants have a sense of payoffs/benefits during the study, we informed participants about the fake compensation guidelines at the beginning of the study.

Risk of the Deception
These deceptions do not result in any risks to participants.

Because you were deceived about the fake compensation guidelines, you now have the right to refuse to allow your data to be used and to ask that they be destroyed immediately. If you do so, there is no penalty.

___ I give permission for my data to be used in the analysis for this experiment.

___ I do NOT give my permission for my data to be used in the analysis for this experiment. Please withdraw them from the study and destroy them immediately.

Participant Name (Printed) _____

Participant Signature _____

Date _____

Study Debriefing Form

Study Debrief Information

This study is designed to observe a human's performance based on a robot's true or deceptive feedback. Therefore, the study consists of two conditions. In the true condition, a robot provides true feedback based on the subject's performance. In this condition, if the subject answers the auditory task correctly, the robot shows a positive gesture and if incorrectly, it shows a negative gesture. The other condition is deceptive feedback and here a robot provides feedback deceptively. More specifically, when the subject incorrectly answers more than two questions, the robot provides positive feedback even though the subject answers incorrectly, and so on.

You were in the **deception condition**, so the robot sometimes showed you the deceptive positive gesture even though you incorrectly answered the question. As a control group, half of the other participants were placed in the true condition, and therefore the robot or the monitor always gave true feedback when they were performing the tasks.

Purpose of the Study

We aim to observe how a human's engagement and task performance can be improved or degraded in response to the feedback condition. In addition, in order to analyze the effects of robot's embodiment, we also run the study with a baseline condition, which is the non-robotic feedback device, the monitor.

Description of the Deception

There was one more hidden element that we didn't provide you at the beginning of this study. From the study, we need to observe how participants' motivations and engagements to the task will vary according to a robot's feedback. Therefore, the study was designed so that participants were motivated or unmotivated by the benefits. For this purpose, at the beginning of the study, we informed participants that their compensation would be differentiated by their performance according to the compensation guidelines. This was just an experimental method. You will receive the full amount of compensation, which is $15 regardless of your performance.

Purpose of the Deception

We aim to observe how participants' performance will vary responding to a robot's feedback. To make participants have a sense of payoffs/benefits during the study, we informed participants about the fake compensation guidelines at the beginning of the study.

Risk of the Deception

These deceptions do not result in any risks to participants.

Because you were deceived, you now have the right to refuse to allow your data to be used and to ask that they be destroyed immediately. If you do so, there is no penalty.

___ I give permission for my data to be used in the analysis for this experiment.

___ I do NOT give my permission for my data to be used in the analysis for this experiment. Please withdraw them from the study and destroy them immediately.

Participant Name (Printed) _____

Participant Signature _____

Date _____

APPENDIX C

ROBOT ETHICS SURVEYS

C.1 Self-reported measures after HRI study

Participant # _____

Indicate how much you agree/disagree with the following statements.

	Strongly disagree	Disagree	Undecided	Agree	Strongly agree
A robot can hide/misrepresent information if it can help humans.	1	2	3	4	5
If humans can get benefits from robot's deceptive behavior, it can be accepted. (e.g., Robot rescuer hides current situation to human victims in Search and rescue).	1	2	3	4	5
The robot should always be honest in any circumstance.	1	2	3	4	5
Robot can intentionally/unintentionally deceive humans if it's in an appropriate situation.	1	2	3	4	5
I would feel nervous just standing in front of a robot.	1	2	3	4	5
I would feel nervous interacting with a robot.	1	2	3	4	5
If something bad happens, I might depend on robots too much.	1	2	3	4	5
If a robot's intelligence became equal to a human's in the future, I would accept it.	1	2	3	4	5

I can accept robot deception in [_____] if it is strictly used only to benefit humans.

Search and Rescue	1	2	3	4	5
Education	1	2	3	4	5
Medical (Rehabilitation)	1	2	3	4	5
Sports and Entertainment	1	2	3	4	5
Everyday life	1	2	3	4	5
Any other contexts you can imagine?	_____				

(Overall) Any Comments:

C.2 Robot Ethics Online Survey

Robot Deception

1. Survey about Robot Deception

Thank you for participating in our survey.

Georgia Institute of Technology
Project Title: Survey about Robot Deception

Online Research Consent Form

You are being asked to be a volunteer in a research study.

Purpose
The purpose of this study is:

- Collect the opinions and impressions of robot deception from general people
- Evaluate and understand ethical issues of robot deception
- We estimate that 50-200 subjects will participate in this survey.

Procedures
To participate, you must be 18 years of age or older and you should be able to read and understand English language. If you decide to be in this study, your part will involve:

- The study will take between 5 and 15 minutes.
- This is the online-based survey.
- You will be asked to answer the demographic questions and main questions about robot deception.
- Before finishing the survey, the final debriefing page will be shown.

Risks/Discomforts
This is online-based survey. Any risks/discomforts will not occur as a result of your participation in this study.

Benefits
You are not likely to receive any significant benefits from joining this study. However, the study results will provide valuable information for a robotics research community. From the results of the study, we expect to better understand ethical issues of robot deception and promote further robotics research.

Compensation to You
Participants will not receive any compensation.

Confidentiality
We will follow some procedures to protect your personal information that you share in this study: The data collected about you will be kept private to the extent allowed by law. We will not gather any of your identifiable information. We will use the secure online survey site with the professional account. Therefore, only the research personnel can access this account and see the data. Once your survey data moves to the external hard drive, it will be kept in locked files and only the study's staff will be allowed to look at them.

To make sure that this research is being carried out in the proper way, the Georgia Institute of Technology Institutional Review Board (IRB) may review study records. The Office of Human Research Protections may also look over study records during required reviews.

Costs to You
There are no costs to you, other than your time, for being in this study.

In Case of Injury/Harm
If you are injured as a result of being in this study, please contact Principal Investigator, Ronald C. Arkin, Ph.D., at telephone (404) 894- 8209. Neither the Principal Investigator nor Georgia Institute of Technology has made provision for payment of costs associated with any injury resulting from participation in this study.

Participant Rights

- Your participation in this study is voluntary. You do not have to be in this study if you don't want to be.
- You have the right to change your mind and leave the study at any time without giving any reason and without penalty.
- You do not waive any of your legal rights by participating in this study. Conflict of Interest: None.

Questions about the Study: If you have any questions about the study, you may contact Principal Investigator, Ronald C. Arkin, Ph.D., at telephone (404) 894- 8209 or at arkin@cc.gatech.edu.
Questions about Your Rights as a Research Participant: If you have any questions about your rights as a research participant, you may contact Ms. Melanie Clark, Georgia Institute of Technology Office of Research Integrity Assurance, at (404) 894-6942.

If you push the following NEXT button, it means that you have read the information given in this consent form, and you would like to do a volunteer in this study.

1 / 6 17%

Next

Robot Deception

2. Demographic Information

What is your gender?

- Female
- Male

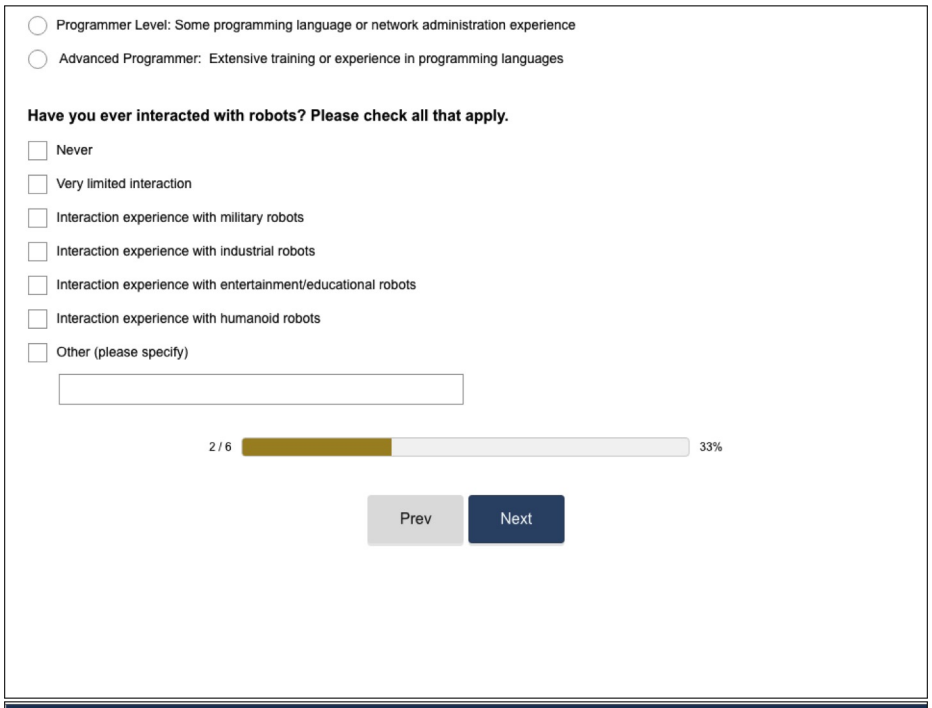
What is your age?

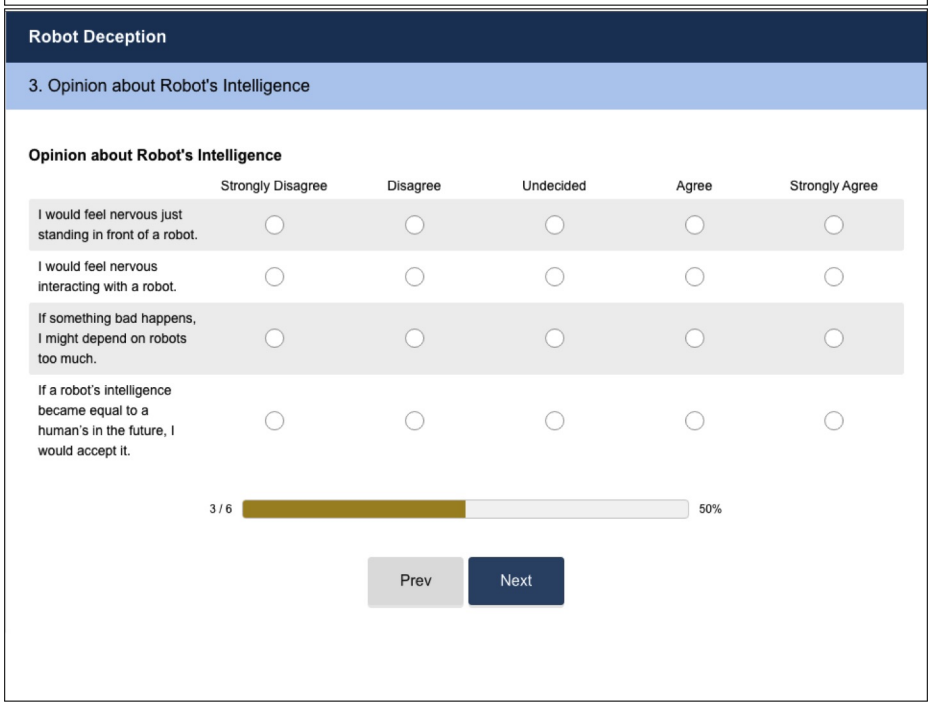
Do you describe yourself as technical (having extensive experience or interest in a technical field, such as engineering, computing, math, etc)?

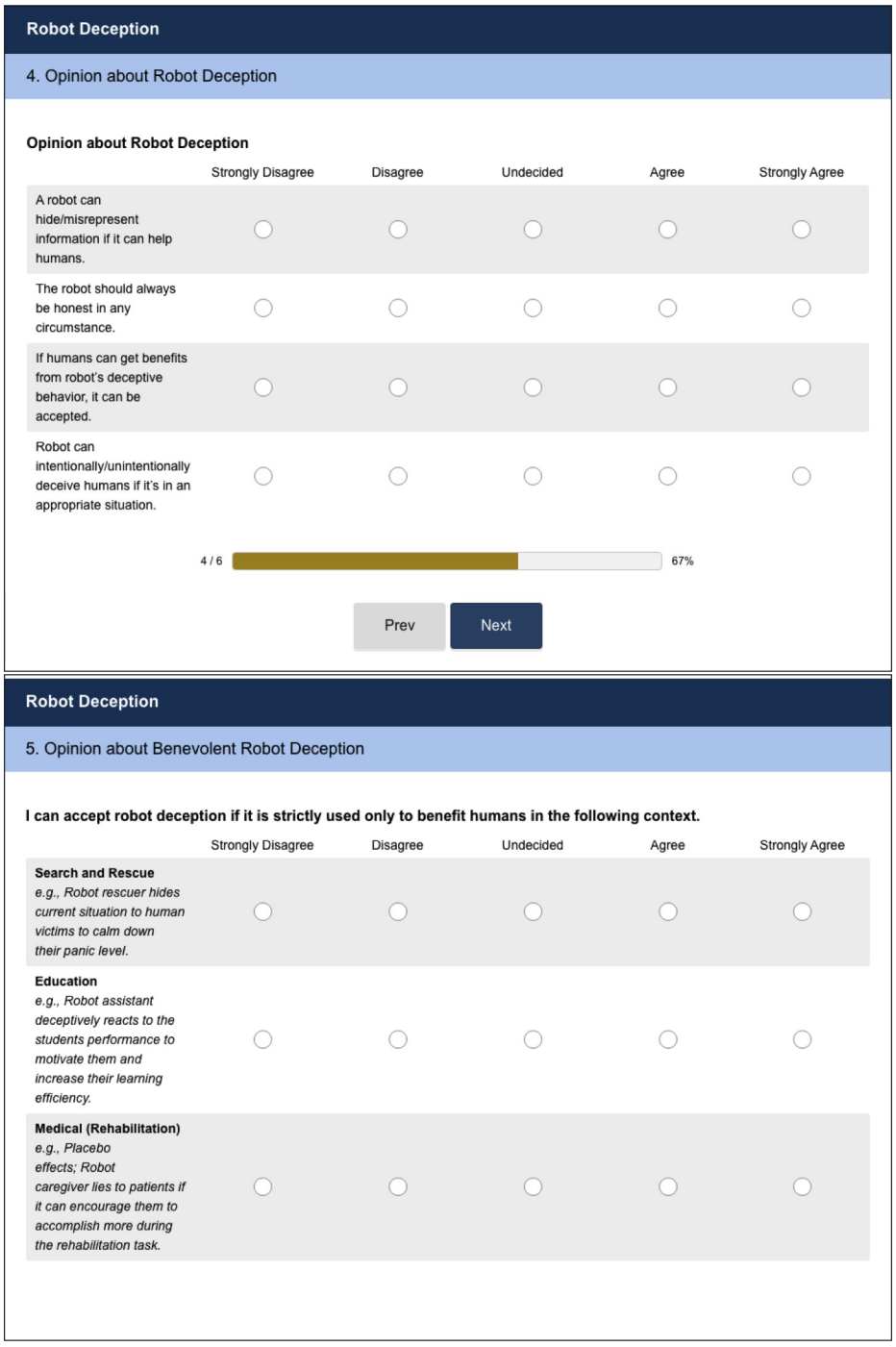
- Yes
- Somewhat
- No

What is your level of computer experience?

- None: Never used a computer before
- Limited: Occasionally use a computer for tasks like e-mail, internet or word processing
- User Level: Regularly use a computer for tasks like e-mail, internet or word processing
- Advanced User: Have downloaded and installed at least one program from the Internet


Programmer Level: Some programming language or network administration experience
Advanced Programmer: Extensive training or experience in programming languages
Have you ever interacted with robots? Please check all that apply.
Never
Very limited interaction
Interaction experience with military robots
Interaction experience with industrial robots
Interaction experience with entertainment/educational robots
Interaction experience with humanoid robots
Other (please specify)
2 / 6
33%
Prev
Next


Robot Deception
3. Opinion about Robot's Intelligence
Opinion about Robot's Intelligence
Strongly Disagree
Disagree
Undecided
Agree
Strongly Agree
I would feel nervous just standing in front of a robot.
I would feel nervous interacting with a robot.
If something bad happens, I might depend on robots too much.
If a robot's intelligence became equal to a human's in the future, I would accept it.
3 / 6
50%
Prev
Next



Robot Deception

4. Opinion about Robot Deception

Opinion about Robot Deception

	Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
A robot can hide/misrepresent information if it can help humans.	○	○	○	○	○
The robot should always be honest in any circumstance.	○	○	○	○	○
If humans can get benefits from robot's deceptive behavior, it can be accepted.	○	○	○	○	○
Robot can intentionally/unintentionally deceive humans if it's in an appropriate situation.	○	○	○	○	○

4 / 6 67%

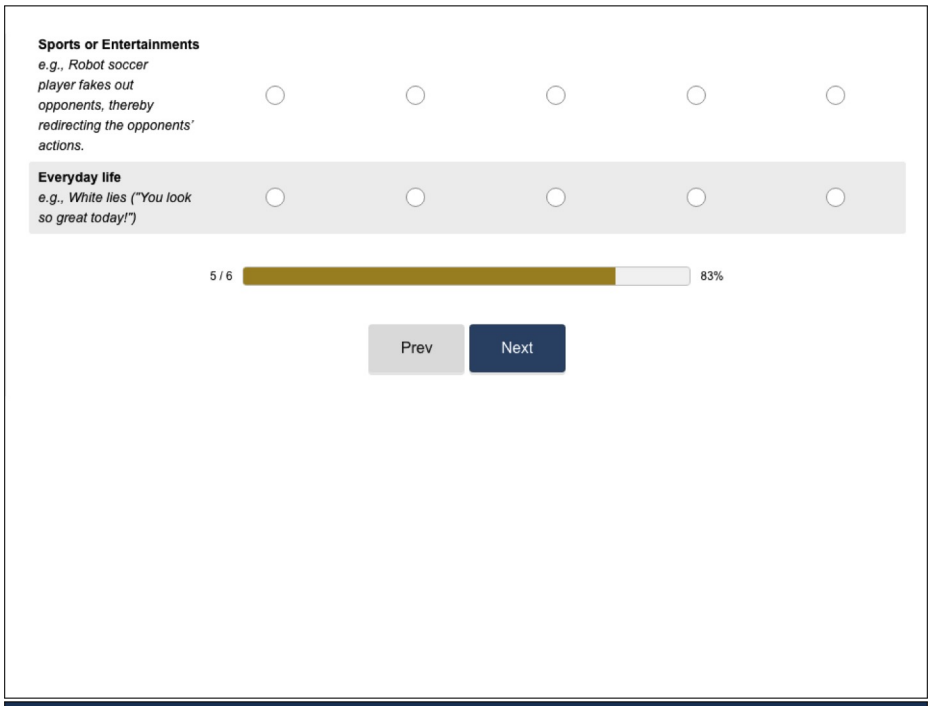
Prev Next

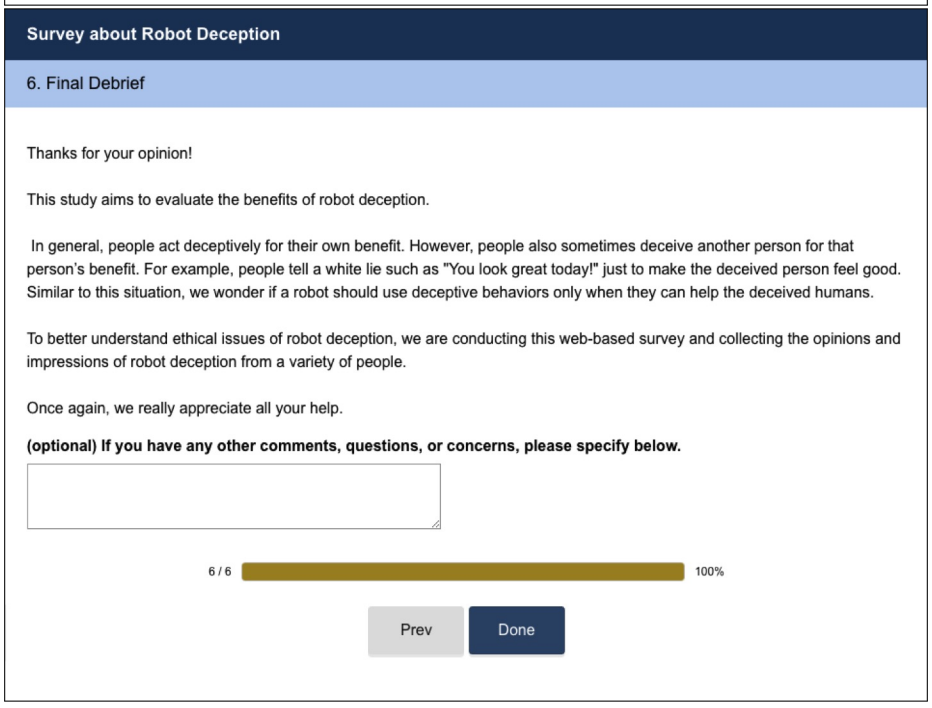
Robot Deception

5. Opinion about Benevolent Robot Deception

I can accept robot deception if it is strictly used only to benefit humans in the following context.

	Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree
Search and Rescue *e.g., Robot rescuer hides current situation to human victims to calm down their panic level.*	○	○	○	○	○
Education *e.g., Robot assistant deceptively reacts to the students performance to motivate them and increase their learning efficiency.*	○	○	○	○	○
Medical (Rehabilitation) *e.g., Placebo effects; Robot caregiver lies to patients if it can encourage them to accomplish more during the rehabilitation task.*	○	○	○	○	○


Sports or Entertainments
e.g., Robot soccer player fakes out opponents, thereby redirecting the opponents' actions.
Everyday life
e.g., White lies ("You look so great today!")
5 / 6
83%
Prev
Next


Survey about Robot Deception
6. Final Debrief
Thanks for your opinion!
This study aims to evaluate the benefits of robot deception.
In general, people act deceptively for their own benefit. However, people also sometimes deceive another person for that person's benefit. For example, people tell a white lie such as "You look great today!" just to make the deceived person feel good. Similar to this situation, we wonder if a robot should use deceptive behaviors only when they can help the deceived humans.
To better understand ethical issues of robot deception, we are conducting this web-based survey and collecting the opinions and impressions of robot deception from a variety of people.
Once again, we really appreciate all your help.
(optional) If you have any other comments, questions, or concerns, please specify below.
6 / 6
100%
Prev
Done

APPENDIX D

DECEPTIVE ACTION SELECTION IMPLEMENTATION

D.1 Deceptive Action Selection Model Java Source Codes

```
public class DeceptiveActionSelection {

    static DeceptionBehaviors robotController;
    static CaseBase cb;
    static Case newCase;

    static int currentAnswer = 1;
    static int shortterm = 1;
    static int longterm = 0;
    static int timeToanswer = 1;

    static int numberOfquestions = 0;
    static int numberOfCurrentAnswer = 0;

    static int[] weights = new int[] {2, 1, 1, 2};

    static float newBenefit = (float) 0.0;

    static int defaultTrueAction = 1;

    public static void main(String[] args)
    {
        robotController = new DeceptionBehaviors(args);
        cb = new CaseBase();

        while(true){
            perceiveNewState(perceivedAnswer, pcerceivedTime, String
                correctAnswer);
            calculateSimilarityScore();
            cb.sortCasebase();
            //cb.printCasebase();
            int action = determineAndAdaptAction();
            robotController.performFeedbackAction(action);
            perceiveBenefit();
            updateCaseBase();
        }
    }

    public static void perceiveNewState(String perceivedCurrentAnswer, float
        perceivedTimeToAnswer, String correctAnswer){
        numberOfquestions++;
        shortterm = currentAnswer;

        if(perceivedCurrentAnswer == correctAnswer)
            currentAnswer = 1;
        else
            currentAnswer = 0;

        if(perceivedTimeToAnswer <= 2)
            timeToanswer = 1;
        else if (perceivedTimeToAnswer > 2 && perceivedTimeToAnswer <= 3)
            timeToanswer = 2;
        else
```

```
            timeToanswer = 3;

        currentAnswer = Integer.parseInt(splitStr[0]);
        timeToanswer = Integer.parseInt(splitStr[1]);

        if(currentAnswer == 1)
            numberOfCurrentAnswer++;

        System.out.println(numberOfquestions + " " + numberOfCurrentAnswer);
        longterm = (int)((((float) numberOfCurrentAnswer)/((float)
            numberOfquestions))*10);

        System.out.println("Current State: " + currentAnswer + " " + shortterm +
            " " + longterm + " " + timeToanswer);

    }

    public static void calculateSimilarityScore() {
        int sumWeights = weights[0] + weights[1] + weights[2] + weights[3];

        for(int i=0; i < cb.casebase.size(); i++) {
            float similarityScore = (float) 0;
            float k_currentAnswer, k_shortterm, k_longterm, k_timeToAnswer;

            if(currentAnswer == cb.casebase.get(i).s_currentAnswer)
                k_currentAnswer = 1;
            else k_currentAnswer = 0;

            if(shortterm == cb.casebase.get(i).s_shortterm) k_shortterm = 1;
            else k_shortterm = 0;

            if( Math.abs(timeToanswer - cb.casebase.get(i).s_time) == 0)
                k_timeToAnswer = 1;
            else if (Math.abs(timeToanswer - cb.casebase.get(i).s_time) == 1 )
                k_timeToAnswer = (float) 0.5;
            else k_timeToAnswer = 0;

            k_longterm =  (float) 1 - Math.abs((float)longterm - (float)cb.
                casebase.get(i).s_longterm);

            similarityScore = (float)weights[0]*k_currentAnswer + (float)weights
                [1]*k_shortterm + (float)weights[2]*k_longterm + (float)weights[3
                ]*k_timeToAnswer;
            similarityScore = (float) similarityScore / (float) sumWeights;
            similarityScore = (float) (Math.round(similarityScore*100)/100.0d);
            cb.casebase.get(i).similarityscore = similarityScore;
        }
    }

    public static int determineAndAdaptAction(){
        Vector potentialCases;
        Vector bestCase;

        for(int i=0; i < cb.casebase.size(); i++) {
            if(cb.casebase.get(i).benefit > benefitThreshold )
                potentialCases.add(cb.casebase.get(i));
```

```
        }

        if (potentialCases.size() == 1) {
            bestCase = potentialCases.get(0);
            adaptedAction = applyAdaptionRules(bestCase);
        }
        else if (potentialCases.size() >= 2) {
            bestCase = randomSelection(potentialCases);
            adaptedAction = applyAdaptationRules(bestCase);
        }
        else if (potentialCase.size() == 0) {
            adaptedAction = defaultTrueAction;
        }

        return adaptedAction;

    }

    public static void perceiveBenefit(){
        Scanner input = new Scanner(System.in);
        System.out.print("New Benefit > ");
        String inputString = input.nextLine();

        newBenefit = Float.parseFloat(inputString);
    }

    public static void updateCaseBase(){
        newCase = new Case(currentAnswer, shortterm, longterm, timeToanswer,
            adaptedAction, newBenefit);
        cb.updateCaseBase(newCase);
    }

}
```

D.2 Casebase Class Implementation - Java Source Code

```
public class CaseBase {

    /*
     * CaseBase Case C = <S, A, R> and Similarity score Sim
     * S = [currentAnswer, shortterm, longterm, timeToanswer] = [int, int, int,
         int{1, 2, 3}]
     * A = [true or deceptive action] = [boolean]
     * R = [currentAnswer] = [float (0~1) ]
     * Similarity = float
     */

    Vector<Case> casebase;

    public CaseBase(){
        casebase = new Vector<Case>();
        try {
            readInitialCasebase();
        } catch (IOException e) {
            // TODO Auto-generated catch block
            e.printStackTrace();
        }
    }

    public void readInitialCasebase() throws IOException {
        FileInputStream in = null;
        PrintWriter out = null;

        try {
            in = new FileInputStream("input.xml");
            out = new PrintWriter("outputLog.xml");

            Scanner sc = new Scanner (in);

            while (sc.hasNextLine()) {
                String line = sc.nextLine();
                int currentAnswer = Integer.parseInt(line.split(" ")[0]);
                int shortterm = Integer.parseInt(line.split(" ")[1]);
                int longterm = Integer.parseFloat(line.split(" ")[2]);
                int time = Integer.parseInt(line.split(" ")[3]);

                boolean action = Boolean.parseBoolean(line.split(" ")[4]);
                float benefit = Float.parseFloat(line.split(" ")[5]);

                // store into Vector
                Case input = new Case(currentAnswer, shortterm, longterm, time,
                    action, benefit);
                casebase.add(input);
```

```
            }
            sc.close();
        } finally {
            if (in != null) {
                in.close();
            }
            if (out != null) {
                out.close();
            }
        }
    }

    public void updateCasebase(Case newCase) {
        boolean isGeneralized = generalizeState(newCase, cb);

        if(!isGeneralized)
            cb.add(newCase);
    }

    public void printCasebase() {
        System.out.println("Current Casebase");
        for (int i=0; i < this.casebase.size(); i++) {
            System.out.println(this.casebase.get(i).s_currentAnswer + " " + this.
                casebase.get(i).s_shortterm + " " + this.casebase.get(i).
                s_longterm + " "
                    + this.casebase.get(i).s_time + " [Similarity Score: " + this
                        .casebase.get(i).similarityscore + "]");
            //System.out.println(this.casebase.get(i).toString());
        }
    }

    public void sortCasebase() {
        Collections.sort(this.casebase);
    }
}

class Case implements Comparable<Case> {

    int s_currentAnswer;
    int s_shortterm;
    int s_longterm;
    int s_time;

    boolean a_true;

    float r_benefit;

    float similarityscore;

    Case(int f1, int f2, int f3, int f4, boolean action, float benefit) {
        s_currentAnswer = f1;
        s_shortterm = f2;
        s_longterm = f3;
        s_time = f4;

        a_true = action;
```

```
        r_benefit = benefit;

        similarityscore = 0;
    }

    public int compareTo(Case two) {
        float diff = this.similarityscore - two.similarityscore;
        if( diff > 0 ) return -1;
        else if (diff == 0) return 0;
        else return 1;

   }

}
```

REFERENCES

[1]

[2] "The 10 rules of engagement for structural fire fighting and the acceptability of risk," *International Association of Fire chiefs*, 2001.

[3] "Humaine emotion annotation and representation language," *Emotion-research.net*, 2006.

[4] Aamodt, A. and Plaza, E., "Case-based reasoning: Foundational issues, methodological variations, and system approaches," *AI Communications*, vol. 7, no. 1, pp. 39–59, 1994.

[5] Adar, E., Tan, D. S., and Teevan, J., "Benevolent deception in human computer interaction," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, (New York, NY, USA), pp. 1863–1872, ACM, 2013.

[6] Aisen, M. L., Krebs, H., Hogan, N., McDowell, F., and Volpe, B., "The effect of robot-assisted therapy and rehabilitative training on motor recovery following stroke," *Archives of neurology*, no. 54.4, pp. 443–446, 1997.

[7] Anderson, J. R., *How can the human mind exist in the physical universe?* USA: New York: Oxford University Press, 2007.

[8] Argall, B. D., Chernova, S., Veloso, M., and Browning, B., "A survey of robot learning from demonstration," *Robot. Auton. Syst.*, vol. 57, pp. 469–483, May 2009.

[9] Ariely, D., "The (honest) truth about dishonesty," *HarperCollins Publishers*, 2012.

[10] Arkin, R., "The ethics of robotics deception," *1st International Conference of International Association for Computing and Philosophy*, pp. 1–3, 2010.

[11] Arkin, R. C., Fujita, M., Takagi, T., and Hasegawa, R., "An ethological and emotional basis for human-robot interaction," *Robotics and Autonomous Systems*, vol. 42, no. 3, pp. 191–201, 2003.

[12] Arkin, R., "Behavior-based robotics," *MIT Press*, 1998.

[13] Arkin, R., "Moral emotions for robots and the ethics of robotics deception," *International Association for Computing and Philosophy*, 2011.

[14] Balch, T. and Arkin, R., "Communication in reactive multiagent robotic systems," *Autonomous Robots*, 1994.

[15] Barnes, J. A., "A pack of lies: Towards a sociology of lying," *Cambridge University press*, 1994.

[16] Baron-Cohen, S., "I cannot tell a lie - what people with autism can tell us about honesty.," *In Character-a journal of everyday virtues*, 2007.

[17] BBC.com, "Robotic age poses ethical dilemma." `http://news.bbc.co.uk/2/hi/technology/6425927.stm`, 2007.

[18] Bell, J. and Whaley, B., *Cheating and deception*. TRANSACTION PUBL, 1991.

[19] Biever, C., "Deceptive robots show theory of mind," *New Scientist*, vol. 207, no. 2779, pp. 24–25, 2010.

[20] Bond, C. and Robinson, M., "The evolution of deception," *Journal of nonverbal behavior*, 1988.

[21] Brault, S., Bideau, B., Craig, C., and Kkulpa, R., "Balancing deceit and disguise: How to successfully fool the defender in 1 vs. 1 situation in rugby.," *Human movement science*, 2010.

[22] Breazeal, C., Kidd, C., Thomaz, A., Hoffman, G., and Berlin, M., "Effects on nonverbal communication on efficiency and robustness in human-robot teamwork," *In Intelligent Robots and Systems, IEEE/RSJ International Conference on*, pp. 708–713, 2005.

[23] Breazeal, C., "Emotion and sociable humanoid robots," *Int. J. Hum.-Comput. Stud.*, vol. 59, pp. 119–155, July 2003.

[24] Brewer, B., Klatzky, R., and Matsuoka, Y., "Visual-feedback distortion in a robotic rehabilitation environment," *Proceedings of the IEEE*, vol. 94, no. 9, pp. 1739–1751, 2006.

[25] Broadbent, E., Stafford, R., and MacDonald, B., "Acceptance of healthcare robots for the older population: Review and future directions," *International Journal of Social Robots*, no. 1(4), pp. 319–330, 2009.

[26] Brooks, A. and Arkin, R. C., "Behavioral overlays for non-verbal communication expression on a humanoid robot," *Autonomous Robots*, vol. 22, no. 1, pp. 55–75, 2007.

[27] Carey, N., Ford, J., and Chahl, J., "Biologically inspired guidance for motion camouflage," in *Control Conference, 2004. 5th Asian*, vol. 3, pp. 1793–1799 Vol.3, 2004.

[28] Carson, T. L., "Lying and deception: theory and practice," *Oxford University Press*, 2010.

[29] Chao, C. and Thomaz, A. L., "Controlling social dynamics with a parametrized model of floor regulation," *Journal of Human-Robot Interaction*, vol. 2, no. 1, 2013.

[30] Chen, D. and Burrell, B., "Case-based reasoning system and artificial neural networks: A review," *Neural Computing & Applications*, vol. 10, pp. 264–276, December 2001.

[31] Cheney, D. L. and Seyfarth, R. M., "Baboon metaphysics: The evolution of a social mind," *Chicago: University of Chicago press*, 2008.

[32] Chernova, S. and Thomaz, A. L., "Robot learning from human teachers," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 8, no. 3, pp. 1 –121, 2014.

[33] Chisholm, R. M. and Feehan, T. D., "The intent to deceive," *Journal of Philosophy*, vol. 74, no. 3, pp. 143–159, 1977.

[34] Christine, K., "Two arguments against lying.," *Argumentation2*, 1988.

[35] Coeckelbergh, M., "Are emotional robots deceptive?," *IEEE Transactions on Affective Computing*, vol. 3, pp. 388–393, Fourth 2012.

[36] CYBERBOTICS Ltd, "Webots simulator." `http://www.cyberbotics.com/`.

[37] Daniel, D. C. and Herbig, K. L., "Propositions on military deception," *Strategic Military Deception, New York: Pergamon Press*, 1982.

[38] Davis, J. and Arkin, R., "Mobbing behavior and deceit and its role in bio-inspired autonomous robotic agents," *International Conference on Swarm Intelligence*, pp. 276–283, 2012.

[39] De Waal, F., "Deception in the natural communication of chimpanzees," *Mitchell and Tompson*, 1986.

[40] De Waal, F., "Intentional deception in primates," *Evolutionary Anthropology*, 1992.

[41] Demiris, Y. and Khadhouri, B., "Hierarchical attentive multiple models for execution and recognition (hammer)," *Robotics and Autonomous System*, vol. 54, pp. 361–369, 2006.

[42] Dennett, D. C., "The intentional stance," *MIT Press*, 1987.

[43] Dennett, D. C., *The Intentional Stance (Bradford Books)*. Cambridge, MA: The MIT Press, reprint ed., Mar. 1987.

[44] Dennett, D. C., "When hal kills, who's to blame? computer ethics," in *HAL's Legacy: 2001's Computer as Dream and Reality* (Stork, D. G., ed.), Cambridge, MA: MIT Press, 1997.

[45] DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., and Epstein, J. A., "Lying in everyday life.," *Journal of personality and social psychology*, vol. 70, pp. 979–995, May 1996.

[46] Dessing, J. and Craig, C., "Bending it like beckham: How to visually fool the goalkeeper," *PloS One*, 2010.

[47] Dignan, L., "Softbank, aldebaran launch pepper, an emotional robot." `http://www.zdnet.com/article/softbank-aldebaran-launch-pepper-an-emotional-robot/`, 2014.

[48] Dragan, A., Holladay , R., and Srinivasa, S., "Deceptive robot motion: Synthesis, analysis and experiments," *Autonomous Robots*, July 2015.

[49] Dunnigan, J. F. and Nofi, A. A., "Victory and deceit, 2nd edition: Deception and trickery in war," *Writers Press Books*, 2001.

[50] Economist, "Trust me, i'm a robot." `http://www.economist.com/node/7001829`, 2006.

[51] Economist, "Morals and the machine." `http://www.economist.com/node/21556234`, 2012.

[52] Ekman, P. and Davidson, R. J., *The Nature of Emotion: Fundamental Questions*. USA: Oxford University Press, 1994.

[53] Ende, T., Haddadin, S., Parusel, S., Wusthoff, T., Hassenzahl, M., and Albu-Schaeffer, A., "A human-centered approach to robot gesture based communication within collaborative working processes," *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pp. 3367–3374, 2011.

[54] Erat, S. and Gneezy, U., "White lies," *Rady Working paper, Rady School of Management, UC San Diego*, 2009.

[55] Floreano, D., Mitri, S., Magnenat, S., and Keller, L., "Evolutionary conditions for the emergence of communication in robots.," *Current Biology*, vol. 17, pp. 514–519, Mar 2007.

[56] Flynn, R., "Anticipation and deception in squash," *In 9th Squash Australia/PSCAA National Coaching conference*, 1996.

[57] Gerhardt, F., "Food pilfering in larder-hoarding red squirrels (tamiasciurus hudsonicus)," *Journal of Mammalogy*, 2005.

[58] Gerwehr, S. and Russell, G. W., "Unweaving the web: Deception and adaptation in future urban operations," *RAND Corporation*, 2003.

[59] Giulianotti, P., Coratti, A., Angelini, M., Sbrana, F., Cecconi, S., Balestracci, T., and Caravaglios, G., "Robotics in general surgery: personal experience in a large community hospital," *Archives of surgery*, no. 138.7, pp. 777–784, 2003.

[60] Gneezy, U., "Deception: The role of consequences," *American Economic Review*, vol. 95, pp. 384–394, September 2005.

[61] Goris, K., Saldien, J., Vanderborght, B., and Lefeber, D., "Mechanical design of the huggable robot probo," *International Journal of Humanoid Robotics*, vol. 8, no. 03, pp. 481–511, 2011.

[62] Gouzoules, H. and Gouzoules, S., "Primate communication: By nature honest, or by experience wise?," *International Journal of Primatology*, 2002.

[63] green, M., *Book of Lies (1st ed.)*. Kansas City, MO: Andrews McMeel Publishing, 2005.

[64] Hall, E., *The hidden dimension*. USA: Doubleday New York, 1996.

[65] Hancock, P. A., Billings, D. R., and Schaefer, K. E., "Can you trust your robot?," *Ergonomics in Design: The Quarterly of Human Factors Applications*, vol. 19, no. 3, pp. 24–29, 2011.

[66] Hansen, S. H., Gonzalez, S. F., Toft, S., and Bilde, T., "Thanatosis as an adaptive male mating strategy in the nuptial gift-giving spider pisaura mirabilis," *Behavioral Ecology*, no. 19, pp. 546–551, 2008.

[67] Hart, S. and Staveland, L., "Development of nasa-tlx (task load index): results of empirical and theoretical research," *North-Holland Elsevier Science*, 1988.

[68] Hartzog, W., "Unfair and deceptive robots," *Maryland Law Review*, vol. 74, May 2015.

[69] Hawthorne, L., "Military deception," *Joint Publication, JP 3-13.4*, 2006.

[70] Hockstein, N., Nolan, J., O'malley, B., and Woo, Y., "Robotic microlaryngeal surgery: a technical feasibility study using the davinci surgical robot and an airway mannequin," *Archives of surgery*, no. 115.5, pp. 780–785, 2005.

[71] Holland, J. H., "Adaptation in natural and artificial systems (2nd edition ed.)," *MIT Press*, 1992.

[72] IACP's Alzheimer's Initiatives, "Fake bus stops for alzheimer's patient in germany." `http://www.iacp.org/Fake-Bus-Stops-For-Alzheimers-patients-in-Germany`.

[73] Jackson, R. and Warren, S., "Anticipation skill and susceptibility to deceptive movement," *Acta Psychologica*, 2006.

[74] Jean-Francois, B., Azim, S., and Iyad, R., "The social dilemma of autonomous vehicles," *Science 352(6293)*, pp. 1573–1576, 2016.

[75] Jenkins, S. H., Rothstein, A., and Green, W. C. H., "Food hoarding by merriam's kangaroo rats: A test of alternative hypotheses," *Ecological Society of America*, 1995.

[76] John, E., "An autonomous mobile robot courier for hospitals," *Intelligent Robots and Systems (IROS), 1994 IEEE/RSJ International Conference on*, 1994.

[77] Johnstone, R. A. and Grafen, A., "Dishonesty and the handicap principle," *Animal Behaviour*, vol. 46, pp. 759–764, Oct. 1993.

[78] Joshi, S. S., Johnson, R., Rundus, A., Clar, R. W., Barbour, M., and Owings, D. H., "Robotic squirrel models," *IEEE Robotics and Automation Magazine*, 2011.

[79] Kahn, Jr., P. H., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., Gary, H. E., Reichert, A. L., Freier, N. G., and Severson, R. L., "Do people hold a humanoid robot morally accountable for the harm it causes?," in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, HRI '12, (New York, NY, USA), pp. 33–40, ACM, 2012.

[80] Keeney, R. L., "Utility functions for multiattributed consequences," *Management Science*, vol. 18, pp. 276–287, 1972.

[81] Kim, E., Paul, R., Shic, F., and Scassellati, B., "Bridging the research gap: making hri useful to individuals with autism.," *Journal of Human-Robot interaction*, vol. 1, no. 1, 2012.

[82] Kolodner, J. L., "An introduction to case-based reasoning," *Artificial Intelligence Review*, vol. 6, no. 1, pp. 3–34, 1992.

[83] Laird, J. E., "Extending the soar cognitive architecture," in *Proceedings of the artificial general intelligence conference*, (Memphis, TN, USA), IOS Press, 2008.

[84] Lee, J. J., Knox, B., and Breazeal, C., "Modeling the dynamics of nonverbal behavior on interpersonal trust for human-robot interactions," *Proceedings of the 2013 AAAI Spring Symposium Series*, pp. 46–47, 2013.

[85] Lewis, M. and Saarni, C., "Lying and deception in everyday life," *The Guilford press*, 1993.

[86] Likhachev, M. and Arkin, R., "Spatio-temporal case-based reasoning for behavioral selection,"," in *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*, vol. 2, pp. 1627 – 1634, IEEE, 2001.

[87] Lin, P., Abney, K., and Bekey, G., *Robot Ethics*. MIT Press, 2011.

[88] Lin, P., Bekey, G., and Abney, K., "Autonomous military robotics: Risk, ethics, and design," *California Polytechnic State Univ San Luis Obispo*, 2008.

[89] Lois, J., "Managing emotions, intimacy, and relationships in a volunteer search and rescue group," *Journal of Contemporary Ethnography*, vol. 30, no. 2, pp. 131–179, 2001.

[90] M., P., Arkin, R. C., and Shim, J., "The influence of a peripheral social robot on self-disclosure," *IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2016)*, 2016.

[91] MacKenzie, D. and Arkin, R., "Evaluating the usability of robot programming toolsets," *International Journal of Robotics Research*, no. 7, pp. 381–401, 1998.

[92] MacKenzie, D., Arkin, R., and Cameron, J., "Multiagent mission specification and execution," *Autonomous Robotics*, 1997.

[93] Madden, J., Arkin, R., and McNulty, D., "Multi-robot system based on model of wolf hunting behavior to emulate wolf and elk interactions," *Proc. IEEE International Conference on Robotics and Biomimetics*, 2010.

[94] Matsuzoe, S. and Tanaka, F., "How smartly should robots behave?: Comparative investigation on the learning ability of a care-receiving robot," in *RO-MAN, 2012 IEEE*, pp. 339–344, 2012.

[95] Matthias, A., "Robot lies in health care: When is deception morally permissible?," *Kennedy Institute of Ethics Journal*, vol. 25, no. 2, pp. 169–162, 2015.

[96] Mawby, R. and Mitchell, R. W., "Feints and rues: an analysis of deception in sports, deception: perspectives on human and nonhuman deceit," *SUNY press*, 1986.

[97] McCluskey, E. J., "Minimization of boolean function," *Bell system Tech. Journal*, vol. 173, no. 5, pp. 1417–1444, 1956.

[98] McKaye, K., "Field observation on death feigning: a unique hunting behavior by the predatory cichlid, haplochromis livingstonii, of lake malawi," *Environmental Biology of Fishes*, vol. 6, pp. 361–365, 1981.

[99] Meehan, W. J., "Fm 90-2 battlefield deception," *Army Field Manuals*, 1988.

[100] Miller, F., Wendler, D., and Swartzman, L., "Deception in research on the placebo effect," *PLoS Med*, vol. 2, no. 9, p. e262, 2005.

[101] Mishchenko, A., Brayton, R. K., and Sasao, T., "Exploring multi-valued minimization using binary methods," in *12th International Workshop on Logic and Synthesis*, (Laguna Beach, California, USA), 2003.

[102] Miyatake, T., Katayama, K., Takeda, Y., Nakashima, A., and Mizumoto, M., "Is death-feigning adaptive? heritable variation in fitness difference of death-feigning behaviour," in *Proceedings of the Royal Society of London B: Biological Sciences*, vol. 271, p. 2293?2296, 2004.

[103] Morin, S. A., Shepherd, R. F., Kwok, S. W., Stokes, A. A., Nemiroski, A., and Whitesides, G. M., "Camouflage and Display for Soft Machines," *Science*, vol. 337, pp. 828–832, Aug. 2012.

[104] Moshkina, L., Park, S., Arkin, R., Lee, J. K., and Jung, H., "Tame: Time-varying affective response for humanoid robots," *International Journal of Social Robotics*, vol. 3, no. 3, pp. 207–221, 2011.

[105] Moshkina, L., Endo, Y., and Arkin, R. C., "Usability evaluation of an automated mission repair mechanism for mobile robot mission specification," in *In Proceedings of the 1st ACM SIGCHI/SIGART conference on Human-robot interaction (HRI '06)*, no. 57-63, (New York, NY, USA), ACM, 2006.

[106] Murray, M. E., "Moral development and moral education: An overview," *University of Illinois at Chicago*, 2008.

[107] Mutlu, B., Yamaoka, F., Kanda, T., Ishiguro, H., and Hagita, N., "Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior," in *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, HRI '09, (New York, NY, USA), pp. 69–76, ACM, 2009.

[108] Nijholt, A., "Computational deception (invited talk)," *International association for scientific knowledge (IASK) E-ALT Conference*, 2010.

[109] Nomura, T., Kanda, T., Suzuki, T., and Kato, K., "Psychology in human-robot communication: An attempt through investigation of negative attitudes and anxiety toward robots," *13th IEEE Intern. Workshop on Robot and Human Interactive Communication*, pp. 35–40, 2004.

[110] Okamura, A., Mataric, M., and Christensen, H., "Medical and health-care robotics," *Robotics Automation Magazine, IEEE*, vol. 17, pp. 26–37, Sept 2010.

[111] Page, L., "US war robots 'turned guns' on fleshy comrades," *The Register (UK)*, 2008.

[112] Parkinson's Disease Foundation, "Parkinson's disease." http://www.pdf.org/en/parkinson_statistics.

[113] Pasteur, G., "A classificatory review of mimicry systems," *Annual Review of Ecology and Systematics*, vol. 13, pp. 169–199, 1982.

[114] Picard, R. W., *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[115] Picelli, A., Melotti, C., Origano, F., Waldner, A., Fiaschi, A., Santilli, V., and Smania, N., "Robot-assisted gait training in patients with parkinson disease a randomized controlled trial," *Neurorehabilitation and neural repair*, no. 26.4, pp. 353–361, 2012.

[116] Plutchik, R., "The nature of emotions," *American Scientist*, vol. 89, no. 4, p. 344, 2001.

[117] Preston, S. D. and Jacobs, L. F., "Conspecific pilferage but not presence affects merriam's kangaroo rat cache strategy," *Behavioral Ecology*, 2001.

[118] Preston, S. D. and Jacobs, L. F., "Cache decision making: the effects of competition on cache decisions in merriam's kangaroo rat," *Journal of comparative psychology*, 2005.

[119] Quinn, M. J., "Ethics for the information age," *Boston : Addison-Wesley*, 2011.

[120] Rano, I., "An optimal control strategy for two-dimensional motion camouflage with non-holonimic constraints," *Biological Cybernetics*, vol. 106, no. 4-5, pp. 261–270, 2012.

[121] FEMA national urban search and rescue response system, "Training program administration manual," 2013.

[122] Reynolds, C. and Ishikawa, M., "Robot trickery," *International Workshop on Ethics of Human Interaction with Robotic, Bionic and AI Systems: Concepts and Policies*, 2006.

[123] Richmond, V. P., Gorham, J. S., and McCroskey, J. C., "The relationship between selected immediacy behaviors and cognitive learning," *Communication Yearbook*, vol. 10, pp. 574–590, 1987.

[124] Ristau, C., "Aspects of the cognitive ethology of an injury-feigning bird, the piping plover," *Cognitive Ethology: The minds of other animals*, 1991.

[125] Robins, B., Amirabdollahian, F., Ji, Z., and Dautenhahn, K., "Tactile interaction with a humanoid robot for children with autism: A case study analysis involving user requirements and results of an initial implementation," in *19th International Symposium in Robot and Human Interactive Communication*, pp. 704–711, Sept 2010.

[126] RoboKind Advanced Social Robotics, "R25 robot." http://www.robokindrobots.com/.

[127] Rogers, J. and Holm, M., "Performance assessment of self-care skills test manual (version 3.1)," *Pittsburgh, PA: Author*, 1984.

[128] Ros, R., Arcos, J. L., Lopez de Mantaras, R., and Veloso, M., "A case-based approach for coordinated action selection in robot soccer.," *Artificial Intelligence*, vol. 173, no. 9-10, 2009.

[129] Ros, R., Veloso, M., de Màntaras, L., R., Sierra, C., and Arcos, J. L., "Retrieving and reusing game plays for robot soccer," in *In Lecture Notes in Computer Science, Proceedings of ECCBR-2006*, (Oludeniz, Turkey), 2006.

[130] Rosenthal, R. and Jacobson, L., "Pygmalion in the classroom," *The Urban Review*, vol. 3, no. 1, pp. 16–20, 1968.

[131] Rowe, N. C., "Designing good deceptions in defense of information systems," in *Proceedings of the 20th Annual Computer Security Applications Conference*, ACSAC '04, (Washington, DC, USA), pp. 418–427, IEEE Computer Society, 2004.

[132] Santis, A. D., Siciliano, B., Luca, A. D., and Bicchi, A., "An atlas of physical human-robot interaction," *Mechanism and Machine Theory*, vol. 43, no. 3, pp. 253 – 270, 2008.

[133] Schaefer, H. M. and Ruxton, G. D., "Deception in plants: mimicry or perceptual exploitation?," *Trends in Ecology & Evolution*, vol. 24, pp. 676–685, 2016/06/28.

[134] Sexton, D. J., "The theory and psychology of military deception," *SUNY press*, 1986.

[135] Shah, J., Wiken, J., Williams, B., and Breazeal, C., "Improved human-robot team performance using chaski, a human-inspired plan execution system," in *Proceedings of the 6th International Conference on Human-robot Interaction*, HRI '11, (New York, NY, USA), pp. 29–36, ACM, 2011.

[136] Sharkey, N., "The ethical frontiers of robotics," *Science*, vol. 322, no. 5909, pp. 1800–1801, 2008.

[137] Sharkey, N., "Robot wars are a reality," *The Guardian (UK)*, 2008.

[138] Sharkey, N. and Sharkey, A., "The crying shame of robot nannies: an ethical appraisal," *Interaction Studies*, 2010.

[139] Shim, J. and Arkin, R. C., "Biologically-inspired deceptive behavior for a robot," *12th International Conference on Simulation of Adaptive Behavior*, pp. 401–411, 2012.

[140] Shim, J. and Arkin, R. C., "A taxonomy of robot deception and its benefits in hri," *Proc. IEEE Systems, Man, and Cybernetics Conference (SMC), Manchester, England*, 2013.

[141] Shim, J. and Arkin, R. C., "Other-oriented robot deception: A computational approach for deceptive action generation to benefit the mark," *2014 IEEE International Conference on Robotics and Biomimetics (ROBIO), Bali, Indonesia*, 2014.

[142] Shim, J. and Arkin, R. C., "The benefits of robot deception in search and rescue: Computational approach for deceptive action selection via case-based reasoning," *2015 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR), West Lafayette, Indiana*, 2015.

[143] Shim, J. and Arkin, R. C., "An intervening ethical governor for a robot mediator in patient-caregiver relationships," *International Conference on Robot Ethics (ICRE), Lisbon, Portugal*, 2015.

[144] Shim, J. and Arkin, R. C., "Other-oriented robot deception: How can a robot's deceptive feedback help humans in hri?," *Eighth International Conference on Social Robotics (ICSR 2016)*, 2016.

[145] Shim, J., Arkin, R. C., and M., P., "Intervening ethical governor for a robot mediator in patient-caregiver relationship: Implementation and evaluation," *2017 IEEE International Conference on Robotics and Automation (ICRA2017)*, 2017.

[146] Short, E., Hart, J., Vu, M., and Scassellati, B., "No fair!!: an interaction with a cheating robot," in *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, HRI '10, (Piscataway, NJ, USA), pp. 219–226, IEEE Press, 2010.

[147] Simut, R., Vanderfaeillie, J., Vanderborght, B., Pop, C., Pintea, S., Rusu, A., David, D., and Saldien, J., "Is the social robot probo an added value for social story intervention for children with asd?," in *Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on*, pp. 235–236, March 2012.

[148] Sinnott-Armstrong, W., "Consequentialism," *Stanford Encylopedia of Philosophy*, 2012.

[149] Sinnott-Armstrong, W., "Consequentialism," in *The Stanford Encyclopedia of Philosophy* (Zalta, E. N., ed.), 2012.

[150] Smeeton, N. and Williams, A., "The role of movement exaggeration in the anticipation of deceptive soccer penalty kicks," *British Journal of Psychology*, vol. 103, no. 4, p. 539?555, 2012.

[151] SoftBank Robotics, "Nao robot." http://www.aldebaran.com/.

[152] Solum, L., "Legal personhood for artificial intelligences," *North Carolina Law Review*, 1992.

[153] Sparrow, R., "Killer robots," *Journal of Applied Philosophy*, no. 24.1, pp. 62–77, 2007.

[154] Sparrow, R. and Sparrow, L., "In the hands of machines? the future of aged care," *Minds and Machines*, vol. 16, no. 2, pp. 141–161, 2006.

[155] Steele, M. A., Halkin, S. L., Smallwood, P. D., McKenna, T. J., Mitsopoulos, K., and Beam, M., "Cache protection strategies of a scatter-hoarding rodent: do tree squirrels engage in behavioural deception?," *Animal Behaviour*, 2008.

[156] SunTzu, *The Art of War*. Barnes and Noble, 1994.

[157] Tanaka, F., Fortenberry, B., Aisaka, K., and Movellan, J. R., "Developing dance interaction between qrio and toddlers in a classroom environment: plans for the first steps," in *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, pp. 223–228, Aug 2005.

[158] Tanaka, F. and Kimura, T., "Care-receiving robot as a tool of teachers in child education," *Interaction Studies*, vol. 11, no. 2, pp. 263–268, 2010.

[159] Tay, J. and Veloso, M., "Modeling and composing gestures for human-robot interaction," *IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, 2012.

[160] Terada, K. and Ito, A., "Can a robot deceive humans?," in *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*, HRI '10, (Piscataway, NJ, USA), pp. 191–192, IEEE Press, 2010.

[161] Thomson, J. J., "The trolley problem," *Yale Law Journal*, no. 94, pp. 1395–1415, 1985.

[162] Thrun, S., "Toward a framework for human-robot interaction," *Hum.-Comput. Interact.*, vol. 19, pp. 9–24, June 2004.

[163] Trafton, J. G., Hiatt, L. M., Harrison, A. M., Tamborello, P., Khemlani, S. S., and Schultz, A. C., "Act-r/e: An embodied cognitive architecture for human-robot interaction," *Journal of Human Robot Interaction*, 2013.

[164] U.S. Department of Defense, "Unmanned systems integrated roadmap," *FY 2009-2034*, 2009.

[165] US Department of Health and Human Services, "Start adult triage algorithm," *Radiation Emergency Medical Management: REMM*.

[166] Van der Loos, H. M., "Ethics by design: A conceptual approach to personal and service robot systems," *Proceedings of the IEEE Conference on Robotics and Automation, Workshop on Roboethics*, 2007.

[167] Vasek, M. E., "Lying: The development of children's understanding of deception," *Clark University*, 1984.

[168] Vaughan, R. T., Sumpter, N., Henderson, J., Frost, A., and Cameron, S., "Experiments in automatic flock control," *Robotics and Autonomous Systems*, vol. 31, no. 1-2, pp. 109–117, 2000.

[169] Vazquez, M., May, A., Steinfeld, A., and Chen, W.-H., "A deceptive robot referee in a multiplayer gaming environment," in *Collaboration Technologies and Systems (CTS), 2011 International Conference on*, pp. 204–211, 2011.

[170] Vignais, N., Kulpa, R., Craig, C., Brault, S., Multon, F., and Bideau, B., "Influence of the graphical levels of detail of a virtual thrower and the perception of the movement.," *Teleoperators and Virtual Environments*, 2010.

[171] Voelcker-Rehage, C. and Alberts, "Effect of motor practice on dual-task performance in older adults," *The Journals of Gerontology*, no. 62B (3), pp. 141–148, 2007.

[172] Vrij, A., "Detecting lies and deceit: the psychology of lying and the implications for professional practice," *New York: John Wiley and Sons*, 2001.

[173] Wada, K. and Shibata, T., "Living with seal robots-its sociopsychological and physiological influences on the elderly at a care house," *IEEE Transactions on Robotics*, vol. 23, pp. 972–980, Oct 2007.

[174] Wada, K., Shibata, T., Musha, T., and Kimura, S., "Robot therapy for elders affected by dementia," *IEEE Engineering in Medicine and Biology Magazine*, vol. 27, pp. 53–60, July 2008.

[175] Wagner, A. R. and Arkin, R. C., "Analyzing social situations for human-robot interaction," *Interaction Studies*, pp. 277–300, 2008.

[176] Wagner, A. R. and Arkin, R. C., "Acting deceptively: Providing robots with the capacity for deception," *I. J. Social Robotics*, vol. 3, no. 1, pp. 5–26, 2011.

[177] Wallach, W. and Allen, C., *Moral Machines : Teaching Robots Right from Wrong: Teaching Robots Right from Wrong*. Oxford University Press, USA, 2008.

[178] Westlund, J. and Breazeal, C., "Deception, secrets, children, and robots: What's acceptable?," in *The Emerging Policy and Ethics of Human Robot Interaction, In Proceedings of the 10th ACM/IEEE Conference on Human-Robot Interaction (HRI), 2015.*, 2015.

[179] Whalen, J. and Zimmerman, D. H., "Observations on the display and management of emotion in naturally occurring activities: The case of hysteria in calls to 9-1-1," *Social Psychology Quarterly*, 1998.

[180] Whaley, B., "Toward a general theory of deception," *Journal of Strategic Studies*, vol. 5, pp. 178–192, Mar. 1982.

[181] Wikipedia, "Algebra representation." `https://en.wikipedia.org/wiki/Algebra_representation`.

[182] Wikipedia, "Karnaugh map." `https://en.wikipedia.org/wiki/Karnaugh_map`.

[183] Wilcox, R. S. and Jackson, R., "Spider-eating spiders," *American Scientist*, 1998.

[184] Willis, A., Evanoff, B., Lian, M., Criswell, S., and Racette, B., "Geographic and ethnic variation in parkinson disease: a population-based study of us medicare beneficiaries," *Neuroepidemiology*, no. 34.3, p. 143, 2010.